

分 类 号_____

学号 D201477734_____

学校代码 10487

密级_____

华中科技大学

博士学位论文

云数据中心的能效计量与系统优化

学位申请人： 姜炜祥

学科专业： 计算机系统结构

指导教师： 金海 教授

答辩日期： 年 月 日

**A Dissertation Submitted in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy in Engineering**

**Energy Accounting and System Optimization
in Cloud Datacenter**

Ph.D. Candidate : Weixiang Jiang
Major : Computer Architecture
Supervisor : Prof. Hai Jin

Huazhong University of Science & Technology

Wuhan 430074, P. R. China

May, 2019

独创性声明

本人声明所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除文中已标明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ，在 ____ 年解密后适用本授权书。
本论文属于

不保密 。

(请在以上方框内打“√”)

学位论文作者签名：

日期： 年 月 日

指导教师签名：

日期： 年 月 日

摘要

数据中心是云计算的底层支撑设施。近年来在云计算的带动下，数据中心的规模不断扩张并在全球得到广泛部署。与此同时数据中心的能耗急剧增长，一方面增加了数据中心的运营成本，另一方面大量的碳排放使数据中心受到来自政府及环保组织的压力与日俱增。在激烈的云计算市场的竞争和日益受到关注的环保压力中，如何有效地控制和降低数据中心的能耗成本以及碳排放已经引起工业界和学术界的广泛关注。

能耗计量是数据中心能耗管理的基石。合理的能耗计量方法可以帮助数据中心分析能耗成本，设计合理的计价策略，并为能效优化提供依据。数据中心的能耗主要由两部分组成：服务器等 IT 设施和制冷系统等非 IT 设施。为了提高 IT 设施的利用率，数据中心通常采用虚拟化技术管理 IT 资源，因此虚拟机是数据中心管理 IT 设施与资源的最基础单元。然而虚拟机作为物理硬件资源的逻辑抽象，无法直接通过硬件测量其能耗。此外，虚拟机之间的资源竞争会导致虚拟机出现能耗变化，而当前的研究中所采用的资源-能耗映射模型无法有效地刻画资源竞争对能耗的影响。另一方面，非 IT 设施也占了数据中心能耗很大比例，而且非 IT 设施的能耗与效率受到 IT 设施负载的影响很大。目前对于非 IT 设施的能耗计量的研究尚不完善。**此外在测量 IT 设施与非 IT 设施能耗关系时发现，数据中心实际运行过程中 IT 负载变化大，而且远远低于设计峰值，导致配套的非 IT 设施利用率和能效低下。**

针对上述问题，基于实际数据中心的能耗测量与分析，并结合经济学博弈理论设计了高效合理的能耗计量方法。此外将 IT 能耗的动态变化和非 IT 设施的能效优化结合，设计实现了高效的制冷系统。具体包括以下三个方面的内容：

在 IT 虚拟机层能耗计量方面，针对解决虚拟机资源竞争导致的能耗变化难以量化的难题，采用经济学博弈理论设计了一种基于宏观准确性和公平性的虚拟机能耗计量评价标准。依据提出的公平性能耗计量标准，将虚拟机的运行状态和经济学方法中的夏普利值方法结合，设计了一种具有公平保证的虚拟机能耗计量方法。**同时针对夏普利值方法输入数量众多的难题，设计了通过局部输入数据获取全局输入数**

据的线性估测方法。通过实验验证，提出的线性估测方法的平均误差小于 6%，最大误差为 11.7%。

在非 IT 设施能耗计量方面，设计了一种基于夏普利值的轻量级计量方法。夏普利值的计算复杂度高达 $O(2^N)$ ，而一个非 IT 设施通常服务成千上万的 IT 设备，因此夏普利值方法难以直接应用到非 IT 能耗的计量。通过测量并分析真实数据中心非 IT 设施的能耗，并基于非 IT 设施能耗特征进行推导证明，设计了一种和夏普利值方法等价但是复杂度仅为 $O(N)$ 的方法。结合实际应用场景并通过实验测量发现，该方法的平均相对误差小于 4%，最大相对误差仅为 6.97%。

在感知 IT 负载的非 IT 设施能效优化方面，设计了一种混合水冷系统。数据中心 IT 负载在时间和空间上存在不均衡的现象。数据中心利用率不高导致制冷系统的利用率低下，存在大量冗余的制冷能力，产生能耗浪费。目前水冷系统由于技术限制，采用的都是中心化的制冷控制方法，导致水冷系统在应对局部热点时制冷效率低下。混合水冷系统将水冷和半导体制冷片相结合，并根据数据中心负载变化设计了细粒度的制冷控制方法。实验结果表明，混合水冷系统在 CPU 制冷能效比上达到了 1.04~1.05 的 PUE。

综上所述，分别面向 IT 层面和非 IT 层面为数据中心设计了全面的、细粒度的能耗计量与分析方法，并结合 IT 负载设计新的制冷系统和控制策略，提高了数据中心的制冷系统能效。

关键词： 数据中心，能耗管理，成本优化，制冷控制，公平性，虚拟机，夏普利值

Abstract

The datacenter is the basic support facility for cloud computing. In recent years, thanks to the promotion of cloud computing, the scale of the datacenter has expanded and deployed widely all over the world. In the meantime, the energy consumption of the datacenter has increased sharply as well. On the one hand, the rising energy consumption has increased the operating costs of datacenters. On the other hand, datacenters are under pressure from governments and environmental organizations because of a large amount of carbon emissions. Under the background of the fierce market competition of cloud computing and the increasing environmental pressure, how to reduce the energy cost of datacenters and carbon emissions has already caused widely concern from industry and academia.

Energy accounting is the bedrock of the datacenter energy management. A reasonable energy metering method is conducive for the datacenter to analyze energy consumption costs, come up with a reasonable pricing strategy and provide the basis for energy efficiency optimization. In a datacenter, the energy consumption is mainly generated by IT and non-IT facilities. For the purpose of improving utilization, datacenters usually use virtualization technology to manage IT resources, which makes virtual machines the basic unit of resource management of datacenters. However, as a logical abstraction of physical hardware resources, virtual machines are not able to measure its energy consumption directly through hardware. Moreover, another problem is that resource competition between virtual machines will lead to changes in energy consumption of virtual machines, while the resource-to-energy mapping model used in the current research cannot effectively reflect the impact of resource competition on energy consumption. On the other hand, non-IT facilities also account for a large proportion of datacenter energy consumption, and their energy consumption and efficiency are greatly affected by the load of IT facilities. The research on energy metering and optimization in this part is still not enough currently. In addition, it

is found that the IT load changes greatly during the actual operation of the datacenter, and is far below the design peak, resulting in low utilization and low energy efficiency of the corresponding non-IT facilities.

To solve the above problems, two efficient and fair energy accounting methods are designed to provide fine-grained IT and non-IT energy of each virtual machine. In addition, a hybrid water cooling system is designed to overcome the drawbacks caused by varied IT load and improve the datacenter energy efficiency. Specifically, it includes the following three aspects:

In order to solve the problem that energy consumption change caused by virtual machine resource competition is difficult to quantify, a new criterion consisting of macro-level accuracy and fairness is designed based on game theory. Based the new criterion, a fair virtual machine energy accounting method is proposed by combining the running state of the virtual machine with the Shapley value method in the economic methodology. The experimental results show that the average error of the proposed method is less than 6% and the maximum error is 11.7%.

To make the Shapley value more suitable for a large scale datacenter, a lightweight energy accounting method based on the Shapley value is developed. The computational complexity of the Shapley value is up to $O(2^N)$, while a non-IT facility typically serves thousands of IT devices. Hence, the Shapley value method is hard to apply directly to the measurement of non-IT energy consumption. By measuring and analyzing the energy consumption of non-IT facilities in real datacenters and deriving on the basis of the energy consumption characteristics of non-IT facilities, a method equivalent to the Shapley value method but with a complexity of only $O(N)$ is designed. The experimental results in real-world deployment show that the average error the proposed method is less than 4% and the maximum error is 6.97%.

To improve the energy usage effectiveness in datacenters, an energy-efficient hybrid water cooling system is designed. The low utilization of servers leads to a large amount of redundant cooling capacity, resulting in wasted energy. Due to technical limitations, cur-

rent water cooling systems generally use centralized cooling control methods, and its energy efficiency also suffers from the utilization imbalance among servers. The new hybrid water cooling system combines water cooling and thermoelectric coolers, and designs a fine-grained cooling control mechanism based on IT load changes to improve the cooling energy efficiency of the datacenter with a partial PUE of 1.04~1.05.

The design of the above two energy accounting methods provides the datacenter a comprehensive and fine-grained energy perspective on the energy consumption, and the IT load-aware hybrid water cooling system improves the energy efficiency of the datacenter.

Key words: Datacenter, Energy Management, Cost Optimization, Cooling Control, Fairness, Virtual Machine, Shapley Value

目 录

摘 要.....	(I)
Abstract	(III)
插图目录	(X)
表格目录	(XI)
1 绪 论	
1.1 研究背景.....	(1)
1.2 国内外研究现状.....	(8)
1.3 研究内容.....	(17)
1.4 论文组织结构	(19)
2 面向 IT 虚拟层的能耗计量方法	
2.1 问题提出.....	(21)
2.2 资源-能耗映射模型的局限性.....	(24)
2.3 基于博弈理论的能耗计量模型	(28)
2.4 动态夏普利值方法.....	(34)
2.5 性能测评.....	(37)
2.6 本章小结.....	(43)
3 面向非 IT 设施层的能耗计量方法	
3.1 问题提出.....	(44)
3.2 非 IT 设施的能耗特性	(46)
3.3 非 IT 能耗计量的定义	(50)
3.4 夏普利值方法的优势与挑战.....	(51)
3.5 轻量级夏普利值能耗计量方法	(53)
3.6 性能测评.....	(58)
3.7 本章小结.....	(62)

4 基于 IT 负载感知的混合水冷系统

4.1 问题提出	(63)
4.2 温水制冷存在的挑战	(66)
4.3 混合水冷系统的设计	(70)
4.4 混合水冷系统的部署框架	(74)
4.5 IT 负载感知的自适应混合水冷控制方法	(75)
4.6 性能测评	(77)
4.7 本章小结	(86)
5 总结与展望	(87)
致 谢	(90)
参考文献	(92)
附录 1 缩略词简表	(103)
附录 2 攻读博士学位期间发表论文	(104)
附录 3 攻读博士学位期间参与的主要科研项目	(105)
附录 4 个人简历	(106)

插图目录

图 1-1 云计算的三种服务形式	2
图 1-2 数据中心构成简图	4
图 1-3 数据中心能耗占比	8
图 1-4 数据中心能耗计量研究现状总结	9
图 1-5 数据中心能耗优化研究现状总结	13
图 1-6 研究内容及创新点总结	18
图 1-7 论文组织结构	19
图 2-1 不同用户使用同一种类型虚拟机时的能耗差异比较	22
图 2-2 虚拟机运行过程中服务器的功耗变化	25
图 2-3 Xeon 服务器能耗变化及虚拟机资源-能耗映射模型的估测	26
图 2-4 Pentium 服务器能耗变化及虚拟机资源-能耗映射模型的估测	26
图 2-5 CPU 超线程技术	27
图 2-6 不同虚拟机集合体的能耗及边际贡献	31
图 2-7 按资源使用比例进行能耗分配的公平性分析	32
图 2-8 基于动态夏普利值方法的虚拟机能耗计量框架	37
图 2-9 虚拟机能耗计量的实验平台	38
图 2-10 同构集合线性评估的误差	40
图 2-11 异构集合线性评估的误差	41
图 2-12 误差累积分布图	41
图 2-13 宏观准确性的比较	42
图 3-1 测量平台结构组成	46
图 3-2 UPS 能耗损和 IT 能耗之间关系	47
图 3-3 空调制冷系统能耗和 IT 能耗之间关系	48
图 3-4 数据中心 IT 能耗负载变化	49
图 3-5 非确定性误差分布	57

华 中 科 技 大 学 博 士 学 位 论 文

图 3-6 二次函数拟合三次函数	58
图 3-7 非确定性误差对 Δ 的影响	60
图 3-8 确定性误差对 Δ 的影响	60
图 3-9 非确定性 + 确定性误差对 Δ 的影响	60
图 3-10 与策略一 UPS 能耗计量的对比	61
图 3-11 与和策略二 UPS 能耗计量的对比	61
图 3-12 与策略三 UPS 能耗计量的对比	61
图 3-13 与策略一 OAC 能耗计量的对比	61
图 3-14 与策略二 OAC 能耗计量的对比	61
图 3-15 与策略三 OAC 能耗计量的对比	61
图 4-1 数据中心水冷系统结构	63
图 4-2 冷水和温水制冷策略的优缺点比较	64
图 4-3 CPU 温度随利用率和水温的变化趋势	67
图 4-4 谷歌数据中心利用率	68
图 4-5 阿里巴巴数据中心利用率	68
图 4-6 同一个水冷系统中不同节点 CPU 温度值 (供水温度 20 摄氏度)	68
图 4-7 热点解决方法对比	69
图 4-8 半导体制冷片	70
图 4-9 混合水冷结构设计	71
图 4-10 混合水冷系统原型	72
图 4-11 CPU 温度随着水温线性增长	73
图 4-12 TEC 的制冷效率	73
图 4-13 混合水冷系统原型	74
图 4-14 数据中心的温度图	76
图 4-15 自适应的混合水冷控制	76
图 4-16 波动型负载中的混合制冷能耗特征	80
图 4-17 不规则型负载中的混合制冷能耗特征	80
图 4-18 常规型负载中的混合制冷能耗特征	81

华 中 科 技 大 学 博 士 学 位 论 文

图 4-19 不同策略的制冷能耗对比	82
图 4-20 水温对波动型负载制冷能耗的影响	83
图 4-21 水温对不规则型负载制冷能耗的影响	83
图 4-22 水温对常规型负载制冷能耗的影响	83
图 4-23 热点比例阈值对波动型负载制冷能耗的影响	84
图 4-24 热点比例阈值对不规则型负载制冷能耗的影响	84
图 4-25 热点比例阈值对常规型负载制冷能耗的影响	84
图 4-26 制冷控制周期对波动型负载制冷能耗的影响	85
图 4-27 制冷控制周期对不规则型负载制冷能耗的影响	85
图 4-28 制冷控制周期对常规型负载制冷能耗的影响	85

表格目录

表 2.1 亚马逊弹性计算云中 16 处理器核心虚拟机每年的电力成本和计算硬件资源成本比较	21
表 2.2 虚拟机能耗计量问题中的主要变量及其含义	28
表 2.3 不同能耗分配策略的比较	29
表 2.4 虚拟机配置及对应的资源-能耗映射模型	39
表 2.5 实验测评中使用的测试程序	40
表 3.1 非 IT 能耗计量问题中的主要变量及其含义	50
表 3.2 三个虚拟机在不同时间内的 IT 能耗 (kW·s)	52
表 3.3 LEAPS 和夏普利值计算时间比较	59
表 4.1 混合水冷实验中其他参数的设置	80

1 緒論

本章首先介绍云计算和数据中心的相关背景，数据中心能耗所面临的挑战以及数据中心具体的组成结构，然后分析国内外针对数据中心能耗的研究现状以及不足之处，接着介绍本文的研究内容和主要贡献，最后介绍论文的组织结构。

1.1 研究背景

1.1.1 云数据中心简介

云计算是一种将计算服务通过互联网传递给普通用户的新型服务，使用户避免了购买以及维护计算机硬件和配套的电力制冷设备等复杂步骤，为用户提供快速的计算服务需求响应。同时，用户可以根据自己的需求动态地调整和购买不同的计算服务，并根据使用时间进行付费。**如图1-1所示，云计算提供的计算服务可以分为三类^[1]：**设施即服务 (Infrastructure-as-a-Service, IaaS)，平台即服务 (Platform-as-a-Service, PaaS) 和软件即服务 (Software-as-a-Service, SaaS)。IaaS 是云计算最基础的服务，主要提供服务器、虚拟机、存储和网络等基础服务。PaaS 主要面向网站和软件开发者，提供实时响应的开发、测试平台和软件管理环境。SaaS 则是通过互联网向用户提供软件应用，由 SaaS 运营商负责维护软件运行环境，而用户只需专注使用软件功能。云计算丰富的功能和便利性使越来越多的用户与公司将业务迁移到云中，云计算的规模也得到了快速的增长。

云计算的稳定服务离不开强大的硬件基础设施的支撑。为了向成千上万的用户同时提供稳定、可靠的云计算服务，云计算提供商需要维护一个由数千甚至数万台服务器组成的 IT 基础设施，同时还需要提供网络接入、稳定电力和精准散热来保证基础设施的可用性。这样一个由电力系统、制冷系统、服务器等硬件实体和虚拟机、应用等软件服务组成的复杂基础设施，称之为数据中心。得益于公共云服务需求的快速增长以及对更加丰富的服务器配置的需求，数据中心产业的发展非常迅速。**根**

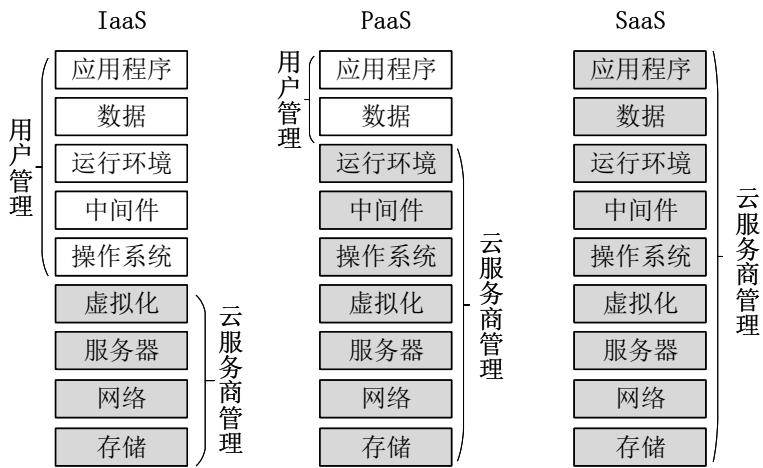


图 1-1 云计算的三种服务形式

据美国调查研究公司 Synergy Research Group 的报告显示^①，2018 年全球数据中心在硬件和软件上的支出增长了 17%，总规模达到了 1500 亿美元，其中公共云基础设施占总数的三分之一以上。此外根据美国第三方市场研究机构 Technavio 的报告分析^②，中国在 2017 年数据中心的市场总规模为 135 亿美元，同比增长 32.4%，并预测截至 2022 年中国数据中心市场将保持 11.55% 的复合年增长率（Compound Annual Growth Rate, CAGR）。

数据中心根据其服务器数量和占地面积可以划分为小型、中型、中大型、大型和超大型数据中心，其中超大型数据中心的服务器数量一般在 5 万至 10 万台。随着大数据、人工智能等技术的发展和产业数字化的转型升级，全球的数据迅猛增长。面对快速增长的数据量，能力强大的数据存储和处理中心的需求将会越来越强。这促使全球数据中心朝着集约化、大型化的趋势发展^[2]，因此超大型数据中心越来越成为一种趋势。根据 Cisco 公司的统计和预测^[3]，全球超大型数据中心在 2016 年底的共有 338 个，其中北美地区的超大型数据中心占全球比例为 48%，亚太地区占比为 30%，而到达 2021 年时超大型数据中心将会增长到 628 个，亚太地区的超大型数据中心数量将超过北美地区，预计到 2020 年，超大规模数据中心客户的服务器购买量将会占

^① Cloud Drives 2018 Spending on Datacenter to \$150 billion, <https://www.srgresearch.com/articles/cloud-drives-2018-spending-data-center-hardware-software-150-billion>

^② Data Center Market in China 2018-2022, <https://www.prnewswire.com/news-releases/data-center-market-in-china-2018-2022-300648785.html>

服务器市场的一半。超大规模数据中心的主要建设者是云计算巨头，如亚马逊、谷歌、微软等^①。

1.1.2 数据中心的能耗挑战

数据中心通常全年 24/7 运行，耗电量巨大。据 2018 年国际科学期刊《自然》发布的报告，全世界数据中心每年消耗 200 太瓦时（TWh）的电量^②。目前信息技术产业消耗全球约 7% 的电力，预计到 2030 年这一比例将上升至 13%^[4]，其中数据中心估计占全球电力消耗的 1.4%。仅在美国，数据中心每年使用超过 900 亿千瓦时的电力，相当于大约 34 个巨型（500 兆瓦）燃煤电厂的发电量。此外根据微软公司的分析报告^[5]，电力设施及其能耗成本已经占到了数据中心总成本的 40%，给数据中心的运营造成巨大压力。

在另一方面，数据中心的能源使用方式和对环境的影响也是数据中心运营商和决策者面临的一个重要问题。根据欧盟委员会的调查报告^[6]，包括数据中心在内的信息和通信技术（Information and Communication Technology, ICT）行业产生的全球二氧化碳排放量在 2017 年已高达全球总排放量的 2%，与航空部门的贡献相当^[7]，其中数据中心的碳排放量增长最快。公众对气候变化和环境影响的看法发生了重大变化，给数据中心的环境政策和社会责任带来了压力。除了优化数据中心本身的能效之外，投资绿色能源发电来减少数据中心碳排放是目前主流数据中心运营商采用的手段之一。亚马逊公司设立了使用 100% 可再生能源为其全球数据中心供电的长期目标^③。目前亚马逊最大的风力发电场—得克萨斯州亚马逊风电场每年为电网增加超过 1,000,000 兆瓦时（MWh）的清洁能源。截至 2018 年 9 月，亚马逊在美国完成了 32 个风能和太阳能项目。这些项目将产生足够的能量，为 283000 个家庭供电。此外谷歌公司于 2017 也宣布了未来实现 100% 绿色能源运营的目标，目前谷歌公司已经实现了超过 30% 的业务运营利用可再生能源进行供电。和国外数据中心发展相比，国内的数据中心运营商在绿色能源投资方面还远远落后^[8]。

^① Hyperscale Data Center Count Jumps to 430, <https://www.srgresearch.com/articles/hyperscale-data-center-count-jumps-430-mark-us-still-accounts-40>

^② How to stop datacentre from gobbling up the world's electricity, <https://www.nature.com/articles/d41586-018-06610-y>

^③ AWS and Sustainability, <https://www.amazon.com/qb>

1.1.3 数据中心的构成

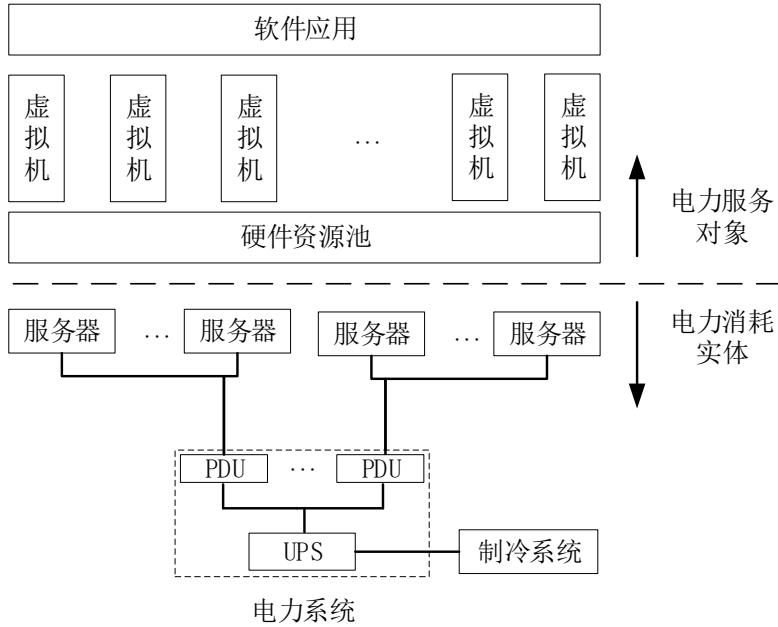


图 1-2 数据中心构成简图

图1-2展示了一个数据中心的简要构成。在软件层面，数据中心通常采用虚拟化技术来进行资源管理。虚拟化技术允许从单个物理硬件系统创建多个模拟环境或专用资源，其中最常见的虚拟化形式是操作系统级虚拟化，其可以在单个硬件上运行多个操作系统。操作系统级虚拟化管理程序直接连接到服务器物理硬件，将一个物理系统拆分为独立的、不同的安全环境，称为虚拟机（Virtual Machine, VM）。每个虚拟机在功能、交互上和一台具有操作系统的服务器相同。虚拟机的优势是可以按照资源需求来进行创建，甚至在运行过程中动态地改变某些资源配置，这种方式有效地避免物理资源的浪费。同时虚拟机还支持在不同物理机之间迁移，方便管理。从前文云计算三种服务形式 IaaS、PaaS、SaaS 的介绍可知，不管哪种云服务，虚拟层都是由云服务提供商进行管理。因此，虚拟机成为数据中心资源管理最基础的单元。

从硬件上来说，数据中心一般由两部分构成：IT（Information Technology）设施和非 IT 设施。IT 设施主要由计算服务器、存储服务器、网络设备等构成。非 IT 设施主要包括电力系统和制冷系统等，为 IT 设施提供稳定电力和合适的运行环境。

数据中心的电力系统主要由不间断电源（Uninterruptible Power Supply, UPS）和

电力分配单元（Power Distribution Unit, PDU）组成。数据中心意外断电可能导致人身伤害，死亡，严重的业务中断或数据丢失。UPS 通常用于保护数据中心内的服务器、电信设备或其他电气设备等硬件。当输入电源或主电源发生故障时，UPS 为这些电气设备提供应急电源。UPS 与一般应急电源系统或备用发电机的不同之处在于，它可以瞬间将电源切换到存储在电池、超级电容器等储电设施中的电力，提供瞬时的输入电源中断保护，从而避免服务器的断电关机和云服务的中断。

电力系统除了为数据中心提供稳定电源之外，本身也消耗很多的电力。电力系统的电力消耗主要由 UPS 的能耗损产生。UPS 的能耗损存在两种形式：按比例损失和固定损失。按比例损失与负载的大小直接相关。相反，无论 UPS 通过多少电流，固定损耗都保持不变。特别的是，当 UPS 负载越低，能耗损占总能耗的比例越大。而数据中心通常会提供冗余的 UPS 系统来保证系统的稳定性和可用性，因此通常单个 UPS 的负载处于较低状态，导致电力系统的能耗损占了数据中心较大的比例。

除了稳定的电力，服务器等设备还需要合适的运行环境。数据中心环境控制是数据中心管理的重点之一，主要功能是将空气的温度，湿度和其他物理参数保持在特定范围内，以便使数据中心内的设备在其整个生命周期内保持最佳性能。当数据中心内的 IT 设备消耗电力时会产生大量热量，用于为 IT 设备供电的 99% 以上的电力都转化为热能^[9]。如果不及时去除多余的热能，室温就会上升，最终导致 IT 设备发生故障甚至火灾。美国采暖制冷和空调工程师协会（American Society of Heating, Refrigerating and Air-conditioning Engineers）发布的数据中心温度管理指南建议数据中心温度应该保持在 68-77°F (20-25°C) 之间^[10]。研究表明，将数据中心保持在 70°F (21°C) 或以下会增加不必要的成本^[11]，而且在相对湿度较高的环境中，过度冷却会使设备暴露在大量水分中，从而促使水蒸气冷凝在电路中的导电细丝上造成短路。虽然较高的机房温度更加节能，但是在较小的机房中，适当的过度冷却可以在冷却系统故障的情况下提供备用热保护。在这种情况下，设备可以继续运行并且在系统因过热而发生损坏之前可以承受几度的温度上升，为应急措施提供反应时间。目前数据中心最常用的制冷手段有三种：空调制冷，室外空气制冷（Outside Air Cooling, OAC）和水冷（又称为液冷）。

空调制冷使用专用的机房空调制冷机（Computer Room Air Conditioner, CRAC）

将机房内的热量传递到外部来将暖空气转换成冷空气。冷空气再通过风扇等空气流通系统进入服务器内部进行冷却。一种常用的空气循环方法是活动地板：CRAC 在房间下方提供冷空气（利用冷空气密度更大，具有更大“重量”的特点），并由风扇从地板向上输送。而后服务器上的风扇再将冷空气吸入服务器内以冷却服务器和其他设备。承载来自服务器废热的暖空气从服务器出来后上升，并且 CRAC 从房间的较高处收集热空气，对空气制冷并将其返回到地板下以重复循环。为了提供更高的效率，一些数据中心设计实现了热通道和冷通道，通过放置隔板（如玻璃隔间）以进一步隔离暖空气和冷空气，其中服务器进气口都面向冷通道，排气口面向热通道。这种类型的设计主要目的是通过一定程度的隔离来最小化热空气和冷空气的混合，提高制冷效率。更复杂的设计还可能涉及机架和天花板之间的隔断设计，以进一步隔离热空气和冷空气。CRAC 也可以将冷却集中在特定通道（而不是整个机房）或甚至特定机架上。这种设计也是为了提供更大的热空气/冷空气隔离，从而提高效率。由于可以将机房和外界进行隔绝保护服务器，空调制冷的优点是可以在任何环境、任何地区建造数据中心。其缺点也很明显，空调制冷非常需要消耗大量电力。尽管设计了很多空调制冷的优化方案，空调制冷的能耗依旧非常高。

室外空气制冷，有时也称自然冷却（Free Cooling），是一种通过使用自然冷却的空气代替机械制冷来降低建筑物或数据中心的空气温度的方法。简单来说，不通过空调制冷，直接将室外的空气引入机房来冷却服务器。在某些季节和时段，许多高纬度地区和较高海拔地区的空气可能比数据中心内的空气要冷得多。通过过滤，加湿并将较冷的主流空气直接引入数据中心，可以减少或消除工业级 CRAC 系统的使用。只有当外部空气温度变得太高，以至于空气无法有效冷却服务器时，才需要 CRAC 系统。因此，安装的 CRAC 系统的工作寿命可以显著延长。CRAC 系统使用的减少也意味着数据中心功耗和服务维修的急剧减少，降低了设施所有者的能源和维护成本。如果外部气温允许持续使用室外空气制冷，则可以完全取消 CRAC 系统。使用这种方法的冷却系统有时被称为空气侧节能器，是一种确保数据中心温度正常的节能方法。这种技术也有其缺陷，因为它可能会使来自户外的污染物和水分进入数据中心。在实际操作中，自然冷却并非完全没有成本，因为需要风扇和其他空气过滤设备，并且这些设备还需要定期维修和维护。

数据中心为了节约土地成本，通常采用结构密集的服务器（如刀片服务器）。由于这种类型的服务器没有预留太多的空气流通空间，给数据中心的冷却系统带来了更大的压力。因此水冷系统被越来越多的数据中心所采用。水冷系统采用液体（如去离子水）作为热媒介。水拥有比空气更好地导热效率，通过管道输入到服务器内部，可以有效地应对结构密集服务器的制冷需求。另一方面，水冷系统可以有效降低数据中心的制冷成本。吸收了服务器热量的水可以输送到室外，通过水蒸发等自然冷却手段降低水温，减少 CRAC 系统的使用。这种方式既具有了室外空气制冷节能的优势，又可以使机房和外界隔绝。水冷却给数据中心带来好处的同时，也带来了新的问题。由于额外的管道，水冷系统可能会限制数据中心设计的灵活性，因为连接到管道的系统不能轻易地拆除重建。电气设备和水的结合也使数据中心维护复杂化。例如，管理员需要定期检查或者事先知道并处理潜在的问题，例如生锈或泄漏。尽管存在固有的挑战，但许多行业专家预测水冷是数据中心制冷不可避免的趋势。

目前数据中心的水冷系统主要针对中央处理器（Central Processing Unit, CPU）、内存等芯片设计，主板等其他电子器件的散热还依赖空调制冷。为了实现全面的水冷，浸没式液冷服务器也逐渐受到业界的关注。浸没式液冷服务器将服务器的所有器件浸泡在特制的低沸点、不导电且无腐蚀性的制冷液中并加以密封，各个元器件产生的热量直接传到制冷液。制冷液的沸点非常低（通常在 60 摄氏度），而且制冷液汽化过程会带走大量热量，而汽化的制冷液通过冷凝再回流到密封机箱内，实现对所有元器件的制冷。然而这种方法缺点也很明显，由于需要密封，服务器的维护非常困难，而且造价较高，目前浸没式服务器还处在探索阶段，没有实现大规模的商用。

图1-3展示了[一般数据中心各类设施所占能耗比例^①](#)，其中服务器、制冷系统、电力系统的占比高达 95% 以上。数据中心能效目前仍旧是一个极其重要的主题，也是数据中心管理的首要任务。这导致出现了许多不同的指标，标准和认证，以确定数据中心的能效。其中最广泛使用的指标是电能使用效率（Power Usage Effectiveness, PUE）。该指标由绿色网格联盟（The Green Grid）创建，被认为是确定数据中心能效的最佳指标之一。PUE 是通过将数据中心的总电量（包含服务器、电力系统、制冷系统、照明等所有设备）除以用于运行其中的计算机基础设施的电量来确定的。PUE

^① CoolIT, <https://www.coolitsystems.com/data-center-cooling-power-myths-busted/>

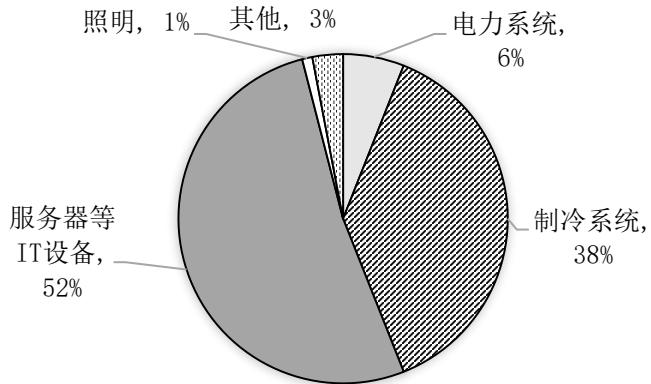


图 1-3 数据中心能耗占比

表示为比率，数值越接近 1，总体能效越高。

美国数据中心调研管理机构 Uptime Institute 对其网络成员进行了调查^①，发现数据中心的平均 PUE 从 2007 年的 2.50 降低到 2014 年的 1.7。亚太地区数据中心远落后于北美先进地区，平均 PUE 在 2014 年至 2018 年仅从 2.04 降至 1.88^②。此外根据我国数据中心信息开放平台公布的数据，我国超大型数据中心和大型数据中心的 PUE 在 2018 年为 1.63 和 1.54^③。在 2019 年，我国工业和信息化部、国家机关事务管理局和国家能源局在联合发布了《关于加强绿色数据中心建设的指导意见》^④，意见中指出大型数据中心的 PUE 要控制在 1.4 以内，可见节能减排一直是数据中心要面对的挑战和追求目标。在数据中心节能减排方面的努力，我国的数据中心目前还落后于美国等先进地区。目前世界能效比最高的数据中心为谷歌公司所运营，其各个数据中心的 PUE 能达到 1.1~1.2。

1.2 国内外研究现状

随着数据中心规模的扩展，能耗问题越来越受到工业界和学术界的关注。目前国内外对数据中心能耗的研究主要可以分为两类：能耗计量和能耗优化。本节将从这两个方面对现有的相关工作进行详细的介绍。

① 2014 Data Center Industry Survey, <https://journal.uptimeinstitute.com/2014-data-center-industry-survey/>

② 腾讯云快讯, <https://cloud.tencent.com/developer/news/379666>

③ 2018 全国数据中心 PUE 情况, <http://www.odcc.org.cn/idcchina>

④ 绿色数据中心建设指导意见, <http://www.miit.gov.cn/n1146295/n1146592/n3917132/n4061768/c6638560/content.html>

1.2.1 数据中心能耗计量

能耗计量是能耗优化的必要条件，可以帮助数据中心了解各种设备的能耗行为模式，并为能耗优化提供依据。如图1-4所示，根据能耗计量的手段不同，数据中心的能耗计量可以分为硬件方法和软件方法。硬件方法主要通过在服务器的电源模块安装传感设备进行监控，如施耐德电气生产的机架智能 PDU^①可以对机架中服务器的能耗进行监控。为了使能耗监控系统更加便捷，SynapSense^②还开发了无线通信的PDU。此外 IBM 公司针对自身的服务器开发了能耗管理系统 Powerexecutive^[12]，其通过在服务器上安装嵌入式的传感模块来实现对每个服务进行能耗监控。上述这种硬件方法的能耗监测粒度受到传感器安装位置的限制，而且成本较高。此外，数据中心对计算资源的使用并不是以服务器为最小粒度，往往是以虚拟机、应用等更细粒度的方式进行计算资源的分配和管理，硬件手段无法有效地对这些对象进行能耗计量和监控。因此低成本且灵活的软件方法越来越受到关注。软件方法主要通过分析资源利用率来进行能耗计量^[13,14]。根据能耗计量的粒度不同，可以将数据中心的能耗计量方法分为四种：部件层、服务器层、应用层和虚拟机层的能耗计量。

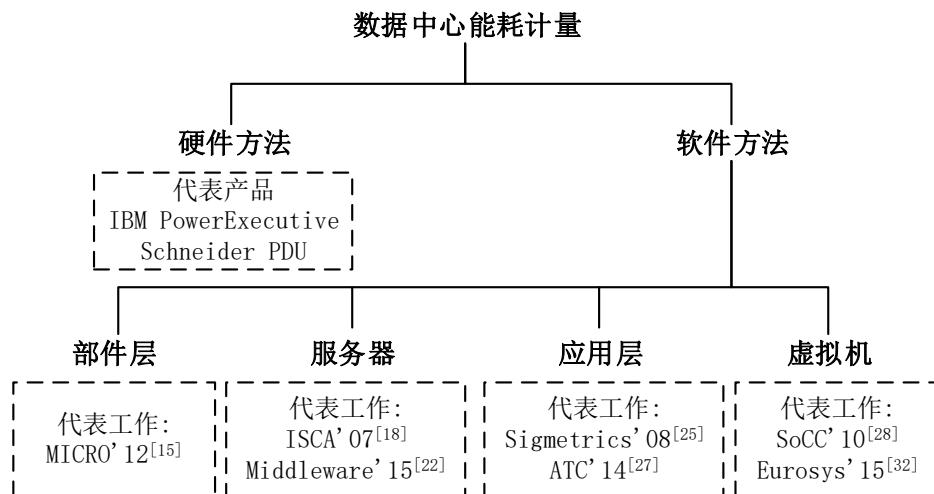


图 1-4 数据中心能耗计量研究现状总结

面向部件层的能耗计量研究：这类工作主要研究如何监控和计量服务器部件的

① Cabinet Power Distribution, <http://www.apc.com/products/family/?id=367>

② Power Monitoring, http://www.synapsense.com/?page_id=22

能耗，从而实现部件级的能耗优化，其中常见的是 CPU 的能耗计量^[15]。运行功耗控制技术 (Running Average Power Limit, RAPL) 由 Intel 公司在 Sandy Bridge 架构 CPU 上推出用于控制 CPU 运行功耗和散热，同时提供了对 CPU 的能耗监控。RAPL 并不是通过硬件手段进行能耗监控和控制，而是通过收集 CPU 不同核心硬件和图像处理单元的使用情况以及 I/O 等，并将这些信息加以不同权重来预测 CPU 的能耗^[16]。由于 CPU 是由多个物理核心构成，各个物理核心的负载、频率等都是独立可控的，因此更加细粒度的能耗信息可以提供更加准确的能耗数据，从而实现更好的能耗优化方法。对此 AMD 公司 Huang 等^[17] 通过结合 CPU 每个引脚的电压、核心频率、核心温度等硬件信息以及 CPU 事件信息，编写了一套分析每个 CPU 核心能耗的硬件逻辑，并以固件的形式部署到了 IBM POWER7+ 架构的系统中，实现了对每个 CPU 核心高精度的能耗计量。

面向服务器的能耗计量：这类工作主要研究服务器的能耗如何监控。由于数据中心服务器数量非常多而且总能耗高，对应配套的电力系统建设成本也非常高昂。因此尽可能地提高数据中心电力设施的利用率可以有效减少浪费。然而数据中心的服务器能耗时高时低，变化非常大，这导致很难确定数据中心应该放置多少的服务器，才能最大化电力设施的利用率同时保证安全性。解决这一问题的有效方法就是对服务器能耗进行监控和计量，谷歌公司的 Fan 等^[18] 利用 CPU 的利用率作为主要参数，并建立利用率和服务器能耗之间的映射关系，从而达到计量服务器能耗的目的，为机柜层、PDU 层以及 UPS 层的电力设施使用提供指导意见。此外，北卡罗来纳大学教堂山分校的 Lim 等^[19] 提出采用多种硬件（包括 CPU、内存、磁盘等）性能计数器，并建立代理线性回归函数来对服务器的能耗进行预测。由于刀片式服务器共享一个刀箱的电力、散热等模块，惠普公司的 Ranganathan 等^[20] 提出了通过跨服务器的资源利用率信息来监控刀箱级的能耗信息。更进一步地，索诺马州立大学的 Rivoire 等^[21] 通过不同的负载测试，比较了五种不同类型的服务器能耗模型，并认为通过操作系统级的资源利用率和 CPU 的性能参数作为能耗模型的输入最为准确。以上工作的方法都基于资源-能耗映射原理，即认为服务器的能耗和各个部件资源利用率直接相关。其中线性的资源-能耗模型由于其构造简单、准确度高而被广泛采用。资源-能耗映射模型的建立需要测量单个服务器的能耗，而对于一个异构且正在运营的数据

中心来说，为每种不同配置的服务器进行能耗测量需要停机安装测量设备，会影响正常的运行。此外，刀片式的服务器共享同一个电源模块而且电源接口特殊，很难找到一种合适的测量设备。对此，加拿大维多利亚大学的 Tang 等^[22,23] 提出了一种非侵入式的服务器能耗估测方法，其通过将服务器进行归类，构建异构的线性能耗模型。特别地，该方法将所有服务器的资源利用率映射到了数据中心整体能耗，而非单个服务器能耗，从而能够在不影响数据中心正常运行，且在只需测得数据中心整体能耗的情况下得出每个服务器的能耗模型，用于能耗计量。加拿大多伦多大学的 Dimitris 等^[24] 还针对部署了数据库处理系统地服务器，通过测量不同操作对 CPU 耗能影响，分析数据库系统服务器的能效。

面向应用层的能耗计量：这类工作主要研究如何计量单个软件应用或者单个计算任务的能耗。软件应用作为电力的最终服务对象，其行为直接决定了服务器能耗的高低。通常软件在开发过程中只注重性能的提升而忽略了能耗。例如不同的库函数可以实现同样的功能，但是最终带来的能耗开销却不一样。为了能让软件在开发阶段就能有效地优化能耗，微软的 Kansal^[25] 等通过研究软件在运行过程中对系统资源的使用行为，设计了一种自动统计软件能耗的工具来指导软件开发过程中对能耗的优化。美国佛罗里达国际大学的 Koller 等^[26] 通过软件本身的特征（如：吞吐率等）而非硬件信息来预测数据中心内软件的能耗。由于一个应用软件或者一个任务在 CPU 中是以线程为基本单位进行调度，因此，线程可以作为计量软件或者任务能耗的一种基本单位。在一个 CPU 核心中，当单个线程和多个线程同时运行时，其能耗状态会有所不同。对此美国威斯康星大学的 Zhai^[27] 等提出一种线程感知的能耗计量模型，通过对 CPU 每个核心上线程数量的判断来对软件或者任务进行能耗计量，提高了能耗预测的准确性。

面向虚拟机的能耗计量：为了提高资源的利用率和管理的高效性，云数据中心通常使用虚拟化技术来管理服务器的资源。因此虚拟机是云数据中心提供用户计算服务的最基础单元，也是进行负载调度与管理基础单元。对虚拟机进行能耗计量，可以有效地指导数据中心在管理过程中对能耗的优化。一般地，虚拟机的能耗计量采用和服务器能耗计量类似的方法，即资源-能耗映射模型。微软研究院的 Kansal^[28] 等通过测量发现 CPU、内存和磁盘是产生虚拟机能耗的主要部件，并利用 Hypervisor 来

获取每个虚拟机对应部件的使用率和 I/O 速度等建立能耗预测模型。然而 Kansal 等发现仅仅利用 CPU 利用率等信息建立的线性能耗模型无法完全准确预测能耗信息。这是由于 CPU 利用率等信息并不能真实反映 CPU 的能耗行为，例如一些负载会在 CPU 负载 100% 的状态的产生 I/O 等待，导致能耗发生变化。而这些细粒度的信息难以直接获取，而且除了 I/O 等待信息之外，类似的细粒度硬件参数数量众多，很难确定哪些参数会影响能耗。因此，Kansal 等采用了不断持续训练并更新能耗模型的策略来减少预测的误差。另一方面，美国乔治亚理工学院的 Krishnan 等^[29]认为仅用 CPU 利用率和内存的利用率无法准确描述虚拟机的能耗，为了提高虚拟机能耗模型预测的准确性，其研究了 CPU 多级缓存的命中率对 CPU 能耗的影响。而 CPU 缓存的命中率会直接影响内存读写操作和内存能耗，因此 Krishnan 等还研究了访存的行为模式对内存系统能耗的影响，从而改进虚拟机的能耗模型。显然，访存的命中率也会影响内存系统和磁盘系统的能耗，纽约州立大学布法罗分校的 Bohra 等^[30]分别将 CPU 缓存命中率与内存读写结合，将内存命中率和磁盘读写结合，提出了由两个线性能耗模型结合的能耗模型。IBM 公司的 Ben-Yehuda 等^[31]为了使用户能在云中部署定制的 Hypervisor，设计了嵌套式的虚拟化系统，使得用户在云提供的虚拟机中再次创建和自由管理虚拟机。这一技术产生了嵌套虚拟机，因此传统的虚拟机能耗计量方法不再适用。对此法国里尔大学的 Colmant 等^[32]提出了一种基于线程的虚拟机能耗计量方法，通过计量嵌套虚拟机中每个线程的能耗来解决嵌套虚拟机能耗的计量问题。[浙江大学的叶可江等^{\[33\]}](#)对近年来数据中心虚拟机能耗计量方法的最近研究成果进行了归纳总结。

1.2.2 数据中心能耗优化

为了应对数据中心能耗增长带来巨大的成本压力，工业界和学术界提出了一系列优化数据中心能耗的方法。如图1-5所示，根据能耗优化的对象不同可以分为 IT 能耗优化和非 IT 能耗优化。其中 IT 能耗的优化主要通过优化 IT 设备的硬件特性和软件负载的调度来实现，而非 IT 能耗主要研究优化电力系统和制冷系统的能耗。

基于硬件特性的能耗优化：这类工作主要利用硬件的特性^[34]来优化能耗，主要的研究对象包括 CPU、内存和磁盘。CPU 的能耗不但与其利用率相关，还和其运行

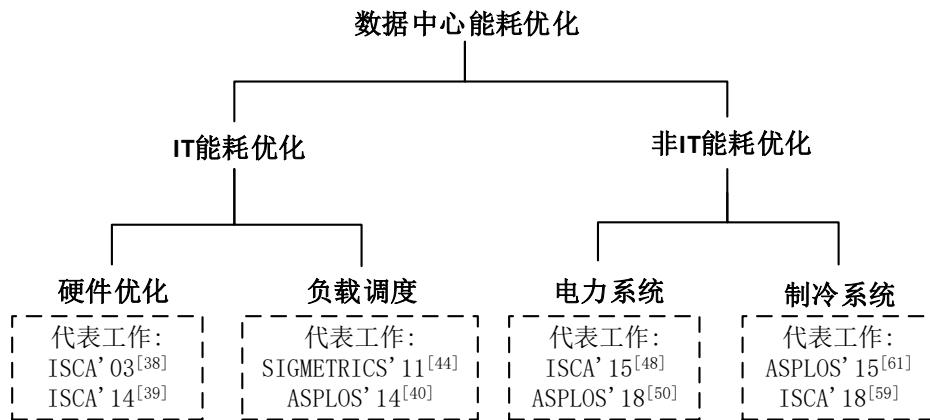


图 1-5 数据中心能耗优化研究现状总结

频率相关，因此通过控制 CPU 的运行频率可以有效控制 CPU 的能耗。但是降低 CPU 的频率会降低性能，因此一般运用在对延迟不敏感的任务上。澳大利亚的斯威本科技大学 Andrew 等^[35] 通过线性模型，将 CPU 能耗和任务的响应时间进行加权结合，从而设计了达到性能和能耗的全局最优的管理策略。伊利诺伊大学厄巴纳-香槟分校的 Michael 等^[36] 认为 CPU 能耗应该与 CPU 温度共同作为优化目标，提出了一种能够同时优化 CPU 能耗与温度的统一方法。和 CPU 类似，内存也可以通过调节工作频率来改变能耗。英特尔公司的 David 等^[37] 首先通过建立内存频率和能耗之间的模型，然后设计了能耗优化算法来降低内存能耗，同时最小化频率变化对内存带宽的影响。目前最常用的磁盘为机械磁盘，该种磁盘的能耗主要由磁盘读写时的转速决定。美国宾夕法尼亚州立大学的 Gurumurthi 等^[38] 研究如何根据负载状态来动态控制磁盘转速来减少能耗，并同时减少对性能的影响。一般来说，通过调节硬件工作状态来节能的办法会影响硬件的性能，因此在实施该种策略时需要考虑对负载任务完成时间、延迟等的影响。美国威斯康星大学麦迪逊分校的 Liu 等^[39] 首先通过测试对性能和功耗状态进行建模，然后基于该模型提出了具有性能保证的能耗优化方法。

基于负载调度能耗优化：新加坡国立大学的 Muthukaruppan 等^[40] 研究了在异构服务器环境中根据不同设备的能耗特性，并利用价格理论来进行负载平衡和任务迁移达到整个服务器集群的能耗最优化。服务器在没有负载时也会产生能耗，但是由于用户任务具有突发性，在服务器空闲时关闭服务器可能会导致无法及时响应用户请求。对此美国密西根大学的 Meisner 等^[41] 提出一种服务器状态保存机制，使服务器

进入一种类似关机的低功耗状态。而当任务来临时，能通过保存的状态迅速恢复服务器的运行状态。这种方法可以大大减少服务器的空闲能耗，而且能最低限度地减少性能损失。犹他大学的 Kshitij 等^[42] 测量发现内存读写速度是众多应用的性能瓶颈，因此提出一种围绕内存性能的任务整合方法，来降低能耗。罗格斯大学的 Eduardo 等^[43] 通过分析存储系统中数据访问的频率优化数据的存放位置，从而减少存储系统的能耗。另外一种能耗优化的负载调度是利用跨域数据中心的不同电价、能效等进行。加州理工大学的 Liu 等^[44] 探讨了跨域数据中心在能耗优化方面的潜力，在提出了两种分布式算法以实现最佳的跨域数据中心负载平衡的同时，还利用不同地区电力的动态定价和负载调度降低数据中心的能耗成本。除了能耗成本，碳排放也是数据中心面临的压力之一，而减少碳排放的有效手段是使用绿色能源（如太阳能、风能等）。绿色能源的发电量受到天气影响不稳定，对此美国田纳西大学的 Zhang 等^[45] 研究如何根据当地天气条件在不同地理位置的数据中心之间动态分配服务请求，以最大限度地利用可再生能源。**华中科技大学的周知等^[46]** 对几种具有代表性的地理分布式云服务进行实证研究，验证电力成本最小化与碳排放最小化之间的冲突，然后综合考虑电力成本，服务质量要求和碳减排目标，利用 Lyapunov 优化技术来设计和分析碳感知控制框架，通过跨域负载调度实现上述三个目标的三向均衡。

面向电力系统的优化：电力系统的优化主要研究如何降低电力设施的建设成本。数据中心在建设时通常根据服务器负载峰值来提供电力设施，而且电力设施一旦建成不容易更改。数据中心的负载高低随着时间而不断发生变化，而且大部分时间都无法达到峰值，这造成了电力设施的极大浪费和不必要的建设成本。因此“超额使用”的策略，即在同一电力系统中增加更多的服务器来提高电力设施的利用率，可以有效减少电力设施的浪费。然而这种方法具有一定的风险，即服务器峰值来临时会造成电力系统的过载问题。加利福尼亚大学圣迭戈分校的 Kontorinis 等^[47] 利用电力系统中的 UPS 来应对电力设施“超额使用”的策略风险。UPS 的作用是在断电时，将服务器的电源供给切换成电池，实现电源不间断从而保证服务器的正常运行。通常断电事件较为罕见，因此电池中的电量可以在“超额使用”的电力设施中作为一种电力补充。在负载较高时，电池放电提高电力系统的电力供给，在负载较低时对电池进行充电。但是电池在充放电的过程中，会产生能耗损失。对此，**西安交通大学的刘**

龙军等^[48]提出了一种将电池和超级电容（Super capacitor）相结合的混合式电力缓冲系统。超级电容具有充放电快且能耗损失低的特点，但是缺点是成本较高，因此刘龙军等提出了根据负载变化的两级充放电控制策略。其中超级电容作为常用的充放电缓冲系统，而电池作为超级电容不足量时的备用充放电缓冲系统，两者结合可以同时降低能耗损失和成本。由于数据中心电力系统是一个多级的电力传输系统，从电网到服务器之间还有 PDU 等电力分配设施。每一级都有各自的负载上限而且负载各异，为了提高各级的电力设施利用率，宾夕法尼亚州立大学的 Wang 等^[49]仿照计算机内存中数据按需换入换出的特点，提出了一种按需分配的电力虚拟化管理系统，根据每个应用程序对能耗的需求，寻找合适的服务器来放置应用程序，同时还结合了 CPU 频率调控、UPS 电池等手段对电力需求和供给进行调整，从而实现整个电力系统的利用率最大化。电力系统的浪费主要是在负载较低时产生，美国密歇根大学的 Hsu 等^[50]通过研究发现数据中心的某些任务负载具有周期变化性，提出了根据任务负载周期变化，将同一时刻低负载和高负载的任务放置在同一级电力设施供电的服务器中，形成负载叠加可以避免该级的电力设施利用率时高时低的现象，提高电力系统利用率的同时，还提高了电力系统的稳定性。

面向制冷系统的优化：温度和散热问题一直以来是计算机发展中绕不开的问题。计算机芯片性能提升主要通过提升晶体管的密度来实现，但是随之而来的散热问题却大大限制了计算机芯片能力的进一步提升。目前已有众多工作研究如何在微体系结构中解决温度问题，如明尼苏达大学 Karen 等^[51]针对 CPU 芯片中不同的物理核心使用程度不同，设计了以降低 CPU 内部温度梯度为目标的 CPU 管理部件，使 CPU 不同物理核心的温度达到均衡，不会因为局部过热导致性能受限。类似地还有针对内存芯片温度管理的相关工作，如伊利诺伊大学厄巴纳-香槟分校的 Aditya 等^[52]在 3D 堆栈内存架构中解决温度问题，来提高系统性能以及美国西北大学的 Majed 等^[53]通过分析管理新型的电阻式随机存取存储器（ReRAM 内存芯片）温度，来提高仿神经计算系统（Neuromorphic Computing Systems）的性能。这类工作的主要目标是解决芯片温度过高导致性能受到限制的问题，而能耗问题并非主要优化目标。数据中心温度管理则更注重在宏观层面进行^[54]，通常以制冷系统的能耗优化作为主要目标。美国杜克大学的 Moore 等^[55]通过对数据中心的温度进行建模并依据模型来判定数据

中心的温度分布（如温度热点等），从而通过任务放置优化数据中心的制冷效率。惠普实验室的 Bash 等^[56] 通过对真实的测量来对一个数据中心内不同服务器的制冷效率来进行排名，并通过任务的合理放置来节省数据中心的制冷能耗。美国罗格斯大学的 Le 等^[57] 通过研究任务放置对制冷能耗的影响，并提出最大化数据中心的运行的温度的任务放置策略，减少制冷系统的工作频率来节省制冷能耗。和电力系统类似，数据中心的制冷需求会随着服务器负载的高低而发生变化，美国密歇根大学的 Skatch 等^[58] 为了减少数据中心制冷系统的建设开销，提出了一种使用相变材料来调整数据中心制冷峰值的方法。相变材料是一种通过状态变化（如固态到液态）来吸热的材料，通过将这些材料部署在服务器中来吸收服务器在高负载时候的热量，然后在服务器负载低时将热量释放，从而使整个数据中心的制冷需求变得平缓，降低制冷系统的峰值负载，减少制冷系统的开销。由于相变材料是靠融化来吸热，只能在超过熔点时工作。为了克服该缺点，Skatch 等进一步提出了通过任务放置来改变服务器内部温度的方法，从而达到主动控制相变材料吸热的目的^[59]。为了减少制冷开销，直接用室外空气进行制冷的方法逐渐兴起，但是室外空气制冷要求数据中心处在温度较低的地区。为了克服室外空气制冷的局限性，美国罗格斯大学的 Manousakis 等^[60] 提出了通过降低性能来减少制冷需求，同时让服务器保持在较高温度的策略，使室外空气制冷的方法可以在较热的地区也能使用。室外空气制冷还存在温度波动大的问题，美国罗格斯大学的 Goiri 等^[61] 研究认为持续的温度变化会引起磁盘系统的不稳定，因此针对采用室外空气制冷的数据中心提出了一种基于温度预测的管理方法，来减少数据中心的温度变化，提高磁盘系统的稳定性。

1.2.3 现有工作的不足之处

虚拟机是数据中心计算资源管理和调度的最基础的方式，因此虚拟机能耗是数据中心能耗管理中最受关注的对象。通过对上述现有研究的总结可以看出，目前对于数据中心各种细粒度能耗主要利用资源-能耗映射模型来进行评估。然而虚拟机作为一种软件层面的硬件抽象，无法直接测量能耗来建立资源利用率和能耗之间的关系，因此其能耗模型通常通过虚拟机状态和服务器能耗建立，其准确性也仅限在服务器层面进行验证。此外，虚拟机之间存在资源竞争会影响虚拟机真实能耗，使得

虚拟机能耗模型准确性更加难以验证，而且能耗模型依赖于各种硬件参数输入，参数选取都是依赖于经验，缺乏理论依据的支撑。而在另一方面，非 IT 能耗占数据中心的比重很大，但是对于非 IT 能耗的细粒度计量，目前尚无相关研究文献。在能耗优化方面，尽管目前 IT 能耗优化的相关工作非常多，而且优化角度全面，但是对于非 IT 能耗优化近几年才开始受到重视，尤其是对新兴的一些非 IT 系统（如水冷系统等）优化工作仍旧不足。目前的能耗优化工作非常依赖于任务负载的调度。在软件系统中，任务负载的调度主要面向性能的提升而非节能，因此性能和节能两个目标之间往往会产生冲突，而且软件系统和制冷系统往往由不同工程人员负责，导致在真实环境中制冷优化的任务放置策略难以部署。

本文认为目前数据中心能耗计量需要解决一个最基本的问题，即虚拟机的 IT 能耗及其非 IT 能耗计量。针对虚拟机的 IT 和非 IT 能耗计量目前面临无法通过实际测量来验证结果的准确性的问题，本文尝试结合经济学中利益分配的原则和传统虚拟机能耗计量中的关键技术结合，分别为虚拟机的 IT 和非 IT 能耗计量设计了具有理论支撑的能耗计量方法方法。而在能效优化方面，本文将尝试从越来越受到关注的水冷系统出发，将水冷系统的优点和数据中心 IT 负载特点结合，帮助数据中心建立一种不影响 IT 负载调度，能自适应 IT 负载变化的高能效制冷系统。

1.3 研究内容

针对上述国内外研究的不足之处，本文设计了新型的数据中心能耗计量方法和能效优化方法。如图1-6所示，针对 IT 能耗和非 IT 能耗这两个构成数据中心能耗成本的因素，本文结合结合经济学中利益分配方法设计了细粒度的公平能耗计量方法；此外在研究 IT 能耗与非 IT 能耗关系时发现，数据中心实际运行中的 IT 能耗远远低于设计峰值，这导致相应的非 IT 设施（如制冷系统）浪费了大量冗余的制冷能力。针对 IT 负载低下导致制冷效率不高的问题，本文将新兴水冷系统的优点和数据中心 IT 负载特点结合，设计了混合水冷系统和对应地制冷控制策略。

(1) 面向 IT 虚拟机层的能耗计量方法。虚拟机的能耗计量一般沿用服务器的能耗计量方法，即采用资源-能耗映射模型，认为能耗的增长和资源的使用率之间存在

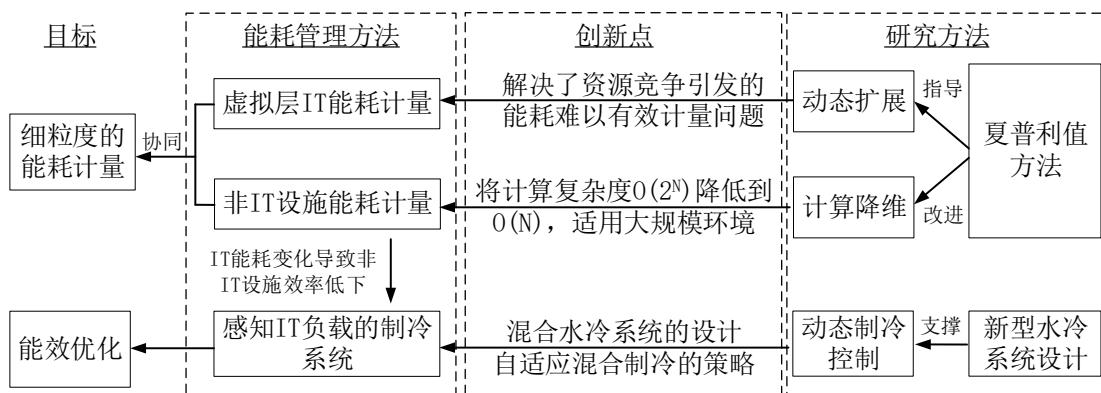


图 1-6 研究内容及创新点总结

线性关系。然而通过实际测量发现，当多个虚拟机共存时，上述的线性关系会发生变化，因此无法用一个统一的能耗模型来描述虚拟机的能耗。上述现象是由虚拟机之间存在资源竞争所导致。资源竞争在 CPU 的使用上尤其显著，CPU 为了提高效率，将单个物理核心通过超线程技术变成两个逻辑独立逻辑核心，而虚拟机正是按照逻辑核心进行资源分配的。这使得虚拟机在 CPU 的使用上产生了一定的竞争关系而使能耗发生了变化。另一方面，虚拟机是一种物理资源的逻辑抽象，无法用硬件对其进行能耗测量和验证，这使虚拟机的能耗计量问题成为一大挑战。本文结合经济学中的利益分配方法，提出了一种新的能耗计量方法。该方法以宏观准确性和公平性作为虚拟机计量的基准，通过引入经济学中的夏普利值方法来进行虚拟机能耗计量。

(2) 面向非 IT 设施层的能耗计量方法。除了 IT 设施外，制冷系统等非 IT 设施的能耗比重占数据中心的 30% 至 50%。不同于 IT 设施可以从物理层面或者逻辑层面进行分割，非 IT 设施由所有的 IT 设施所共享，难以进行有效的划分。因此细粒度的非 IT 能耗计量成为一大挑战。本文结合夏普利值方法，提出了基于公平性的非 IT 能耗计量策略。特别地，夏普利值方法的计算复杂度高达 $O(2^N)$ ，在一个物理机能服务的虚拟机数量有限，但是一个非 IT 设施可能服务于成千上万的虚拟机，这使得夏普利值方法在该场景中难以应用。对此，**本文测量总结了真实数据中心内 IT 能耗与非 IT 设施的能耗关系**，并基于这些能耗特征进行严格的推导证明，得出一种和夏普利值的方法等价，但是复杂度仅为 $O(N)$ 的能耗计量方法，有效地解决了非 IT 能耗的计量问题。

(3) 基于 IT 负载感知的混合水冷系统。本文在测量 IT 能耗与非 IT 设施的能耗关系时发现，数据中心在实际运行中 IT 能耗负载远低于数据中心的设计峰值，导致配套的非 IT 设施利用率低下，存在能耗浪费的现象。本文将结合新兴的水冷技术，设计混合水冷系统和相应的控制策略来克服上述缺点，提高数据中心能效。当前的水冷系统制冷策略非常保守，采用的制冷水水温较低，因此存在大量冗余的制冷能力，而通过提高水温可以有效减少水冷系统的能耗，但是提高水温在面临突发负载时存在制冷失败的风险。另一方面，水冷系统缺乏细粒度的制冷控制手段且数据中心的负载具有不均衡的特点，这使制冷的效率受到局部热点的限制。对此，本文采用温水制冷策略并设计了水冷和半导体制冷片相结合的混合水冷系统以及 IT 负载感知的自适应混合制冷策略，来提高数据中心的制冷能效，并通过真实的系统原型进行了有效性验证。

1.4 论文组织结构

本文围绕数据中心能耗计量和优化两个问题展开研究，分为五章内容，其结构组织如图1-7所示。

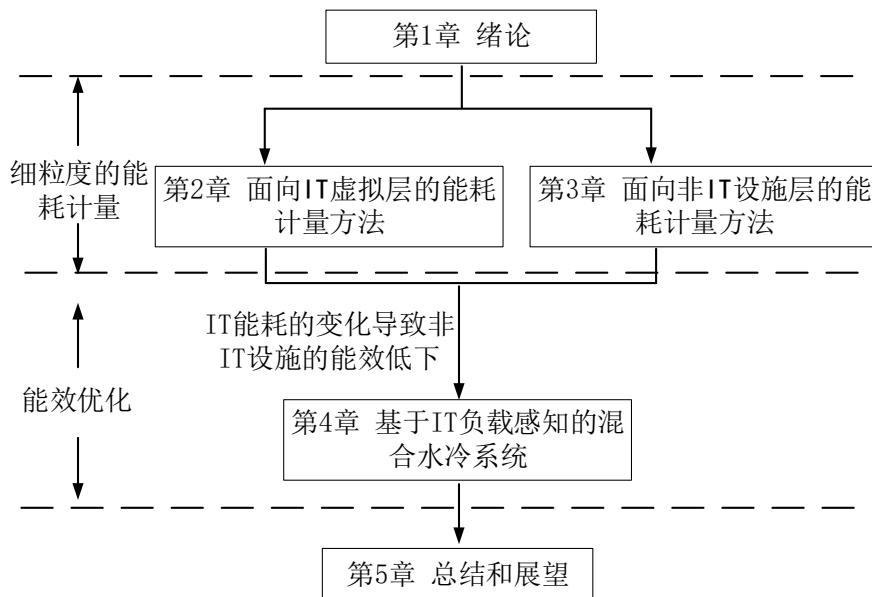


图 1-7 论文组织结构

第 1 章为绪论部分。本章首先对云计算和数据中心的发展进行了基本介绍，同时阐明了数据中心面临的能耗组成及其问题与挑战；之后分析了国内外针对数据中心能耗管理方法研究的现状和最新技术手段，讨论了现有工作中能耗管理方法存在的不足之处和空白；最后阐述了本文的研究内容和主要贡献，并给出了全文的组织结构。

第 2 章为面向 IT 虚拟机层的能耗计量方法。首先介绍能耗对虚拟机运行成本和计价的影响，并通过测量验证现有虚拟机能耗计量方法的不足之处，分析虚拟机资源竞争导致能耗难以准确计量的挑战，然后提出了基于公平性的能耗计量方法。最后根据真实的系统测量，对本文提出的能耗计量方法进行性能评估。

第 3 章为面向非 IT 设施层的能耗计量方法。首先介绍了非 IT 能耗在数据中心能耗的占比以及影响，分析了非 IT 设施由于 IT 设备共享导致能耗难以有效计量的挑战。然后通过对真实数据中心中 IT 能耗和非 IT 能耗的测量，分析了非 IT 设施的能耗特征，并借助该能耗特征对夏普利值进行推导证明，得到一种基于夏普利值方法但是复杂度极低的能耗计量方法。最后分析了误差并通过真实的数据中心能耗数据集验证了本文提出的非 IT 能耗计量方法的准确性和可行性。

第 4 章为基于 IT 负载感知的混合水冷系统。首先描述了数据中心水冷系统的构造和控制原理，同时介绍了温水制冷的策略以及该策略在现有水冷系统中应用时存在的问题。然后提出了一种新型的混合水冷系统来应对温水制冷带来的挑战，并通过分析数据中心在时空负载的差异性，设计了自适应的混合水冷控制方法。最后通过搭建的真实混合水冷系统原型，利用真实的数据中心负载数据集对提出的硬件系统和控制方法进行了性能测试评估，验证了其有效性。

第 5 章总结全文和创新点，并展望数据中心能耗管理的发展方向和可能存在的研究问题。

2 面向 IT 虚拟层的能耗计量方法

数据中心日益增长的能源消耗给云计算运营商带来了巨大的成本压力。为了达到成本与收益的合理平衡，一种可行的途径是根据数据中心众多用户的实际资源和能源消耗进行精准公平计费。然而，虚拟机作为云计算数据中心内物理资源的复用抽象和租用对象，一直以来无法通过硬件手段直接测量其精确能耗，这也使得现有虚拟机能耗估测方法准确性无法直接验证。另一方面，租户虚拟机之间存在相互竞争物理资源的现象，该现象导致的无法通过简单的资源使用率来进行虚拟机的能耗划分。为了量化这种资源竞争对虚拟机能耗的影响，本章通过经济学中的博弈理论，将虚拟机能耗测量问题提炼并转化为成本分配问题，提出了基于公平性的能耗计量评价标准，并结合虚拟机各维度资源和实际系统因素设计了动态夏普利值方法（Dynamic Shapley Value）方法，为数据中心中虚拟机能耗实现公平高效的计量结果。

2.1 问题提出

数据中心作为云计算的底层硬件支撑的设施与系统，其能耗也随着云计算规模的飞速扩展而不断增长。根据国际自然资源保护协会的评估^[62]，数据中心已经成为世界上电力消耗增长最快的设施。快速增长的能耗给数据中心的运维成本造成了巨

表 2.1 亚马逊弹性计算云中 16 处理器核心虚拟机每年的电力成本和计算硬件资源成本比较

虚拟机实例类型	电力成本（美国）	电力成本（德国）	CPU 成本	内存成本	存储成本
通用型	\$100.74	\$193.52	\$310.4	\$80	\$26
计算优化型	\$105.15	\$201.94	\$349	\$40	\$26
访存优化型	\$100.74	\$193.52	\$310.4	\$160	\$26
存储优化型	\$100.74	\$193.52	\$310.4	\$160	\$256

¹ 计算硬件资源的更新周期按 5 年计算。

² 计算硬件资源成本按照亚马逊弹性计算云中对应的 16 核心虚拟机实例计算。能耗成本按照对应的 CPU 设计功耗以及 2015 年美国和德国的平均电力价格计算。

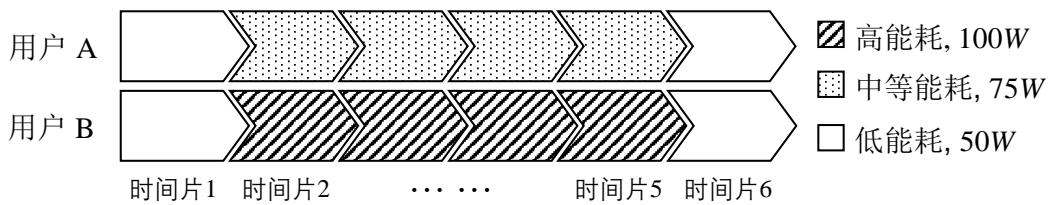


图 2-1 不同用户使用同一种类型虚拟机时的能耗差异比较

大的挑战。为了应对数据中心快速增长的能耗成本，学术界和工业界提出了一系列的数据中心节能方案，如，通过虚拟机实例的整合，关闭利用率较低的服务器，甚至“超用”数据中心的电力设施等。尽管数据中心的能耗本身的成长速度有所放缓，但由于数据中心的规模仍然在扩张，而且电力价格也不断上涨，其能耗成本仍旧不断增长。表 2.1 总结了在亚马逊弹性计算云（Amazon Elastic Computing Cloud, EC2）中，支撑一台 16 处理器核心的虚拟机所需的电力成本^①和计算硬件资源成本^②。很显然，电力成本已经和计算硬件资源的成本相当，而且随着运营时间的延长，电力成本会不断累积，甚至超过计算硬件资源的成本。

尽管能耗已经成为数据中心运营的巨大成本，但是这部分成本并没有在目前的云数据中心虚拟机计价策略有所体现，比如亚马逊弹性计算云^③和阿里云^④只是简单地按照虚拟机的配置和使用时间来进行计价，而隐藏了能耗成本。事实上，这种计价策略会造成对用户收费不公平的现象。如图 2-1 的例子所示，用户 A 和用户 B 在相同的时间内租用了相同配置的虚拟机。然而，在虚拟机使用的过程中，两个用户任务和计算资源使用的行为模式各异，这导致两个用户实际产生了不同的能耗。很显然，图 2-1 例子中的用户 B 在 [时间片 2, 时间片 5] 之间所产生的电力成本比用户 A 多了 33%，然而按照当前的计价模型，用户 A 和用户 B 所支付的费用是一样的，意味着两个用户平摊了数据中心的能耗成本，这显然对用户 A 不公平。因此，在一个多用户的云数据中心中，在计价策略中考虑每个用户实际的能耗成本尤为关键。实际上，类似的能耗计价策略已经在互联网数据中心（Internet Data Center, IDC）业务中得到

^① Global electricity prices by select countries in 2015, <http://www.statista.com/statistics/263492/electricity-prices-in-selected-countries/>

^② Amazon EC2 Instance Types, <https://aws.amazon.com/ec2/instance-types/>

^③ Amazon EC2 Pricing, <https://aws.amazon.com/ec2/pricing/on-demand/>

^④ 阿里云定价, <https://www.aliyun.com/price/>

应用^[63,64]。IDC 是一种托管用户服务器的数据中心，并会针对用户服务器所消耗的电力进行计费。与之不同的是，云数据中心的能耗计价的策略要求能够监控每个虚拟机所消耗的能耗。

然而，虚拟机作为一种计算硬件资源的逻辑抽象，无法通过硬件直接测量出其能耗，因此虚拟机的能耗测量一直是数据中心能耗管理面临的一大难题。**对虚拟机进行能耗建模是目前虚拟机能耗测量的一种通用方法^[29,30]**。能耗模型最初被用来进行单个服务器的能耗估测，其主要是通过测量服务器所产生的能耗，并同时监测物理硬件资源的消耗情况（如 CPU 利用率），建立物理硬件资源的使用率和能耗之间的映射关系^[65]。类似地，通过监测单个虚拟机在单个服务器上运行时，虚拟机对硬件资源的消耗与服务器能耗，并建立两者之间的映射关系，就可以得到虚拟机的能耗模型。然而，该方法存在运用到虚拟机能耗估测时存在缺陷。通常在一个服务器上会同时运行多个虚拟机，而这些虚拟机之间存在物理资源上的竞争（如 CPU 计算单元、寄存器、缓存等）。虚拟机在竞争这些物理资源时，会导致虚拟机对物理资源使用行为的变化，从而影响到各个虚拟机真实产生的能耗。这种资源竞争的现象，会导致虚拟机能耗模型的失效。根据本章后续实际测量发现，虚拟机的能耗模型在多虚拟机共存的场景中，误差高达 25.22% 至 46.15%。

本章针对虚拟机物理资源竞争导致能耗难以有效测量的难点，提出一种基于公平的能耗分解方法。该方法首先利用合作博弈理论，将虚拟机能耗的测量的工程问题抽象为成本分配的经济学问题。成本分配问题的核心就是公平。由于虚拟机是一种计算资源的逻辑抽象，而且通常多个虚拟机共存于一个服务器而且互相影响，真实能耗无法直接测量和验证。因此将服务器的能耗在各个虚拟机之间进行公平分配是一种新型的、有效的能耗计量思路。成本分配问题可以通过夏普利值（Shapley value）方法进行求解。此外，虚拟机的能耗会随着其不同时刻，对物理资源使用的状况发生变化，因此需要实时地、动态地进行估测。然而夏普利值法是一种静态的成本分配策略，对此，本章将夏普利值方法扩展动态夏普利值方法，该方法的核心是利用同一服务器中所有虚拟机对物理资源整体使用率，和服务器能耗之间的映射关系，得到在不同状态下服务器的能耗。该能耗作为夏普利值方法的动态输入，最后得到对应状态下各个虚拟机的能耗。

2.2 资源-能耗映射模型的局限性

本节将具体介绍资源-能耗映射模型估测虚拟机能耗所存在的问题。具体而言，本节通过真实的实验测量，验证多个虚拟机共存时，资源-能耗映射模型的失效问题，然后介绍并分析资源竞争对虚拟机能耗测量所带来的挑战。

2.2.1 实验设置

X86 处理器是目前应用最广泛的服务器处理器，同时在云数据中心中得到了广泛的部署。因此，实验选取了两个不同架构的 X86 处理器（Intel Xeon 和 Intel Pentium）服务器作为实验测量平台。此外，服务器各连接了一个电能计量表，该电能计量表可以记录服务器的实时功率。由于当多个虚拟机共存于服务器时，无法直接测量每个虚拟机的能耗。因此，**本小节首先采用已有的资源-能耗映射模型的构建方法^[28]**，来获取虚拟机的能耗模型。该方法使用“边际能耗变化”来作为虚拟机的真实能耗，即保持其他虚拟机的状态不变的情况下，用一个虚拟机启动或者关闭时，服务器的能耗变化作为该虚拟机的真实能耗。此外，实验在每个服务器上各开启两个配置 1 CPU 核心，512MB 内存，8GB 磁盘以及 Linux 系统的虚拟机，记为 C_VM 和 C_VM'。其中 CPU 是最耗能的部件，因此，为了更清晰地展现实验结果，实验中保持虚拟机其他部件（内存、磁盘读写、网络 I/O 等）不变，只采用了 CPU 利用率作为虚拟机消耗的资源。

2.2.2 资源-能耗映射模型的构建与验证

实验首先验证资源-能耗映射模型在服务器层面的准确性。为了得到该能耗模型，首先在 Intel Xeon 服务器上启动虚拟机 C_VM 和 C_VM'，并在虚拟机中运行测试程序使虚拟机的 CPU 利用率发生变化，同时记录虚拟机的 CPU 利用率和服务器对应的能耗信息。通过记录的信息可以得到如下服务器能耗模型：

$$p' = 9.49u' + 138 \quad (2.1)$$

p' 代表了服务器的能耗， u' 代表了两个虚拟机 CPU 利用率的总和，138 代表了

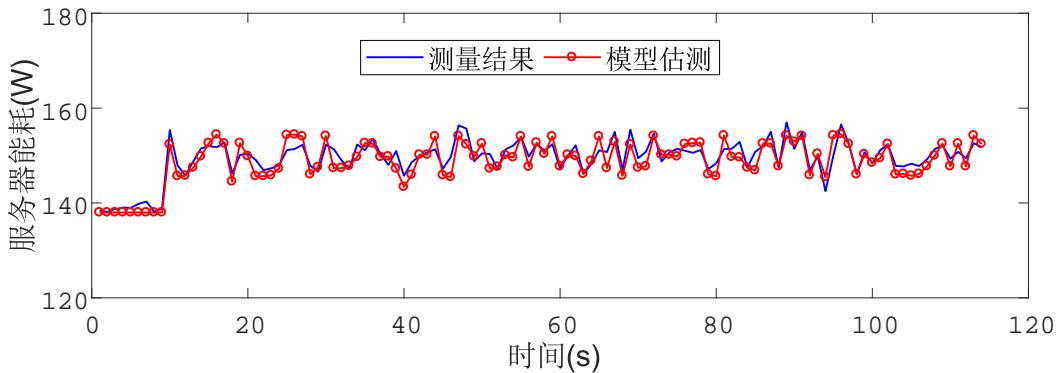


图 2-2 虚拟机运行过程中服务器的功耗变化

服务器的静态能耗^①。图2-2展示了该模型使用虚拟机 C_VM 和 C_VM' 的 CPU 利用率去估测服务器能耗的结果。可以看到，该能耗模型在服务器层的准确性很高，平均误差仅为 2.07%。该观察结果和现有工作^[28] 的测量结果一致。

实验其次验证资源-能耗映射模型在虚拟机层面的准确度。通过在两个虚拟机中运行 Linux Shell 指令 (“scale=6000; 4*a(l)”| bc -l -q) 执行长时间的浮点运算任务，可以将虚拟机的 CPU 利用率保持在 100%，而其他内存等部件的使用率基本为 0。根据资源-能耗映射模型，在运行该浮点运算的过程中，两个虚拟机对资源使用率的状态完全一致，因此虚拟机 C_VM 和 C_VM' 对服务器的能耗贡献应该是相同的。图2-3和图2-4展示了在两个虚拟机中执行该浮点运算任务时，服务器的能耗变化。显然，图中的测量结果和资源-能耗映射模型的估测的结果并不相符。

以图2-3中 Xeon CPU 的服务器作为例子进行具体分析，当两个虚拟机都保持空闲状态时，服务器的能耗为 138W。当在 C_VM 虚拟机中提交浮点运算任务时，服务器的能耗增加到了 151W。因此，根据资源能耗映射模型，虚拟机的能耗模型为：

$$p = 13u \quad (2.2)$$

其中 p 代表了虚拟机的能耗， u 代表了 CPU 的利用率。

可以看出，当浮点运算任务只在一个虚拟机中提交时，资源-能耗映射模型的预测准确度比较高，其平均误差为 0.23%。因此根据该能耗模型，当浮点运算任务在另一个虚拟机上提交时，服务器的能耗应该增加 13W。然而，当同样的浮点运算任务提

^① 静态能耗是指服务器在所有的虚拟机保持空闲状态下的能耗，该能耗是一个稳定值。

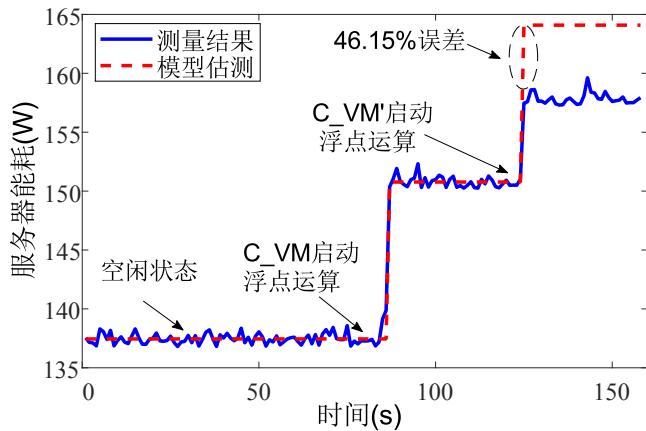


图 2-3 Xeon 服务器能耗变化及虚拟机资源-能耗映射模型的估测

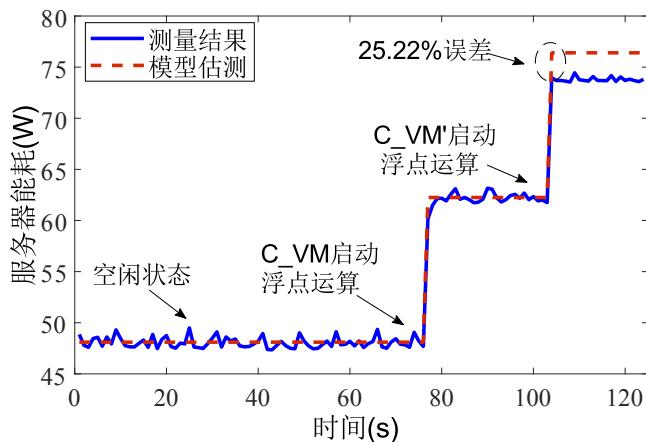


图 2-4 Pentium 服务器能耗变化及虚拟机资源-能耗映射模型的估测

交时，电能计量表的测量结果显示，服务器的能耗只增加了 7W。这与能耗模型的预测结果相比，产生了 46.15% 的相对误差。实验还调整了浮点运算任务在两个虚拟机上执行的顺序，仍然观察到类似的结果：即第一个提交任务的虚拟机使服务器增加了 13W 的能耗，而第二个提交任务的虚拟机使服务器增加了 7W 的能耗。显然，资源能耗映射模型在估测虚拟机的能耗时，并不完全适用，而且当虚拟机的处理器核心数量增加时，能耗估测的绝对误差会进一步增大。

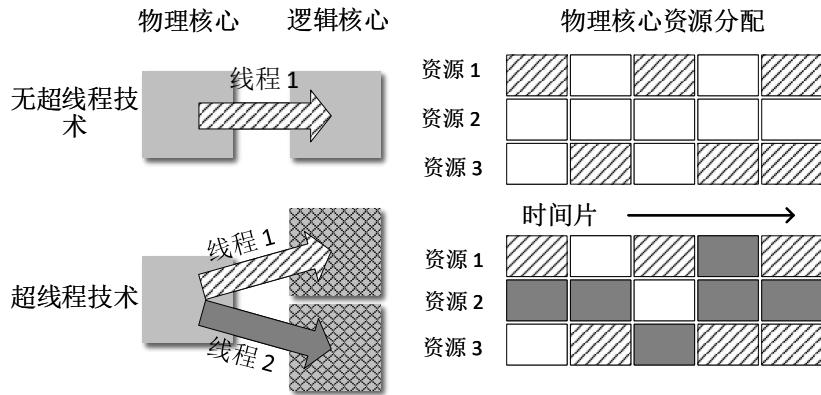


图 2-5 CPU 超线程技术

2.2.3 虚拟机资源使用行为分析

导致资源-能耗映射模型失效的原因是 CPU 广泛采用了一种超线程技术 (Hyper Threading Technology, HTT)。图2-5展示了该技术的原理图。超线程技术通过复用物理核心的方式，来同时执行两个线程的计算任务，从而提高 CPU 的效率。因此，一个物理核心在操作系统中变成了两个可独立执行计算任务的逻辑核心。显然，当两个逻辑核心需要同时使用同一个物理资源（如，寄存器等）时就会产生资源竞争，其中一个逻辑核心必须等待另外一个逻辑核心使用完并释放该物理资源。而虚拟机的处理器核心是按照逻辑核心来进行分配的。因此，实验中两个虚拟机在同一个服务器上执行浮点运算任务时，就会产生类似的资源竞争。在任意一个时刻，事实上只有一个虚拟机在使用该物理资源。服务器能耗的增加或减少，是由服务器中物理资源的真实使用情况所决定的。显然，虚拟机的 CPU 利用率无法判断真实的物理资源使用情况，而操作系统无法捕捉到这种细粒度的资源竞争，这使得资源-能耗映射模型出现了失效的情况。

服务器在开启后，即使处于空闲状态下，也会产生比较稳定的能耗，该能耗称为静态能耗。静态能耗并不是由虚拟机或其他任务运行所产生，而是维持服务器开启后，保持待机状态所必需的。在服务器运行任务或虚拟机时，产生的能耗称为动态能耗。因此，在本节之后的内容中，进行虚拟机能耗以及服务器能耗比较时，均不包括该静态能耗，即服务器能耗仅指动态能耗。关于静态能耗的如何在虚拟机之间分配，将会在后续章节中分析。

表 2.2 虚拟机能耗计量问题中的主要变量及其含义

变量	含义
\mathcal{N}	同一服务器上 n 个虚拟机的集合
\mathcal{S}	虚拟机集合 \mathcal{N} 的子集；一个虚拟机集合体
k	一个虚拟机中部件状态的数量
\mathbf{c}_i	虚拟机 i 的部件状态向量
\mathcal{C}	一个虚拟机集合体的部件状态向量
\mathbf{v}_j	属于同一个同构虚拟机集合体 j 的虚拟机部件状态向量和
\mathbf{w}_j	同构虚拟机集合体 j 的能耗映射向量
Φ_i	虚拟机 i 的能耗
$v(\mathcal{S})$	虚拟机集合体 \mathcal{S} 的能耗
$\Phi_i(\mathcal{C})$	虚拟机集合体状态为 \mathcal{C} 时虚拟机 i 的能耗
$v(\mathcal{S}, \mathcal{C})$	虚拟机集合体状态为 \mathcal{C} 时虚拟机集合体 \mathcal{S} 的能耗

2.3 基于博弈理论的能耗计量模型

本节首先定义虚拟机的能耗计量问题，然后介绍夏普利值的相关背景，以及如何使用夏普利值法计量虚拟机的能耗。

2.3.1 虚拟机能耗计量问题的定义

本章假设在任意一个时刻，一个服务器上运行的虚拟机实例的集合为 \mathcal{N} 。虚拟机能耗计量的目标是确定该时刻任意一个虚拟机 i 对服务器能耗 P （不包含静态能耗）的贡献 Φ_i , $i \in \mathcal{N}$ 。从数学上说，虚拟机能耗计量是建立虚拟机能耗 Φ_i 和服务器能耗 P 之间的函数映射关系。

一般而言，虚拟机能耗计量方法需要满足准确性，即估测出的能耗的 Φ_i 等于虚拟机 i 的真实能耗。然而，由于多个虚拟机共存于一个服务器并且相互影响，因此无法获取虚拟机的真实能耗，也无法验证上述准确性条件。针对此问题，本章提出一种新的虚拟机能耗计量结果的评估方法：

1. 宏观准确性：估测的虚拟机能耗之和等于测量的服务器能耗，即 $\sum_{i \in \mathcal{N}} \Phi_i =$

P 。

2. 公平性：每个虚拟机的所分配到的能耗 Φ_i 是公平的。

表 2.3 不同能耗分配策略的比较

分配策略	C_VM	C_VM'	C_VM + C_VM'	服务器能耗	宏观准确性	公平性
边际增长分配	13 W	7 W	20 W	20 W	✓	✗
资源-能耗映射模型	13 W	13 W	26 W	20 W	✗	✓
理想策略	10 W	10 W	20 W	20 W	✓	✓

为了进一步阐述宏观准确性和公平性，本节使用图2-3中的两个虚拟机（C_VM 和 C_VM'）对不同分配方法进行比较，结果如表2.3所示。

根据对服务器能耗的边际增长贡献进行分配，第一个虚拟机的能耗为 13W，第二个虚拟机的能耗为 7W。显然，虚拟机的能耗之和等于服务器的能耗。这满足了宏观准确性条件。但是两个配置相同、执行任务相同、状态相同的虚拟机全分配到了不同的能耗，违背了公平性条件。

根据资源-能耗映射模型，两个完全相同的虚拟机所分配到的能耗相等，各为 13W。这满足了公平性条件。但是虚拟机的能耗之和不等于服务器的能耗，无法满足宏观准确性条件。

从图2-3可以看出，两个完全相同的虚拟机，对服务器的总能耗贡献为 20W。显然，一种合理的分配方策略是每个虚拟机各分配到 10 W 的能耗。这种分配策略既满足了宏观准确性条件，又满足了公平性条件。但是，当各个虚拟机的配置以及状态均不相同时，很难找到一种同时满足上述两个条件的分配策略。这是因为宏观准确性条件，可以通过测量进行验证，而公平性条件难以量化验证。

2.3.2 基于夏普利值的虚拟机能耗计量

为了有效解决上述虚拟机能耗计量公平性的挑战，本章提出一种基于合作博弈理论的虚拟机能耗计量方法。如何在多个参与者组成的合作博弈中分配利益所得，是合作博弈理论中的经典问题^[66,67]。该问题的核心就是公平性，即如何使得每个参与

者的利益分配是公平可合理的，这与上述的虚拟机能耗计量问题十分吻合。在虚拟机能耗计量问题中，每一个虚拟机就是一个合作博弈的参与者。而同一台服务器上的虚拟机组成了一个合作博弈，服务器的能耗就是虚拟机相互影响，合作所产生的“利益”。虚拟机能耗计量的目标就是如何公平地将该“利益”分配给每一个虚拟机。

夏普利值方法^[68] 在 1953 年被提出并用于解决上述合作博弈的利益分配问题，是合作博弈中被广泛认可的一种公平的利益分配方法，该方法也被广泛应用在计算机资源管理问题中，如网络带宽的分配问题^[69,70]。在夏普利值方法中，将虚拟机的一个子集 $\mathcal{S} \subseteq \mathcal{N}$ 称为一个虚拟机集合体。对于每一个虚拟机集合 \mathcal{S} ，都有一个价值函数 (Worth Function) $v(\mathcal{S})$ ，其表示该虚拟机集合体 \mathcal{S} 所产生的能耗总和。

夏普利值方法之所以被广泛认可，是由于其被证明符合合作博弈理论中的四条关于公平性定义的公理：

公理 2.1：（效率公理 Efficiency）每个合作博弈参与者分配所得之和，必须等于总的利益，即

$$\sum_{i \in \mathcal{N}} \Phi_i = v(\mathcal{N}). \quad (2.3)$$

公理 2.2：（对称公理 Symmetry）如果合作博弈参与者 $i, j \in \mathcal{N}$ 对任意的一个虚拟机集合体的贡献是相同的，则 i 和 j 是对称的，且分配所得相同。即对所有可能的且不包含 i 和 j 虚拟机集合体 $\mathcal{S} \subseteq \mathcal{N}$, $i, j \notin \mathcal{S}$, $v(\mathcal{S} \cup \{i\}) = v(\mathcal{S} \cup \{j\})$ ，则 $\Phi_i = \Phi_j$ 。

公理 2.3：（无效性公理 Dummy）如果一个合作博弈的参与者对所有可能的虚拟机集合体的贡献都为零，则该参与者的所得分配也为零，即对于任意的 $\mathcal{S} \setminus \{i\} \subseteq \mathcal{N}$, $v(\mathcal{S} \cup \{i\}) = v(\mathcal{S})$ ，则 $\Phi_i = 0$ 。

公理 2.4：（可加性公理 Additivity）任何两个独立的合作博弈联合在一起时，那么参与者在组成的新博弈中获得的分配 (Φ_i)，等于在两个独立合作博弈中获得分配 (Φ'_i 和 Φ''_i) 之和，即 $\forall i$, $\Phi'_i + \Phi''_i = \Phi_i$ 。

此外，夏普利值方法具有唯一性，即其被进一步证明是唯一符合上述四条公理的利益分配方法^[68]。具体而言，其数学定义如下：

$$\forall i \in \mathcal{N}, \Phi_i = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{|\mathcal{S}|!(|\mathcal{N}| - |\mathcal{S}| - 1)!}{|\mathcal{N}|!} \cdot [v(\mathcal{S} \cup \{i\}) - v(\mathcal{S})]. \quad (2.4)$$

其中 $|\mathcal{N}|$ 和 $|\mathcal{S}|$ 分别代表了集合 \mathcal{N} 和 \mathcal{S} 的基数。夏普利值的含义可以进行如下解读：假设每个虚拟机都是按先后顺序加入同一个物理机，且在虚拟机 i 加入之前，同一个物理机上存在的虚拟机集合为 \mathcal{S} 。那么 $v(\mathcal{S} \cup \{i\}) - v(\mathcal{S})$ 则代表了虚拟机 i 加入物理机所带来的能耗边际增长。根据排列顺序，子集 \mathcal{S} 中的虚拟机加入物理机的顺序有 $|\mathcal{S}|!$ 种，而在虚拟机 i 之后加入的其他虚拟机有 $(|\mathcal{N}_j| - |\mathcal{S}| - 1)!$ 种顺序。分母 $|\mathcal{N}|!$ 代表将所有虚拟机的可能序列取平均值，整个分数项代表了一个特殊的加权值。因此，夏普利值是将虚拟机 i 在加入不同虚拟机子集所产生的不同边际贡献进行加权平均。

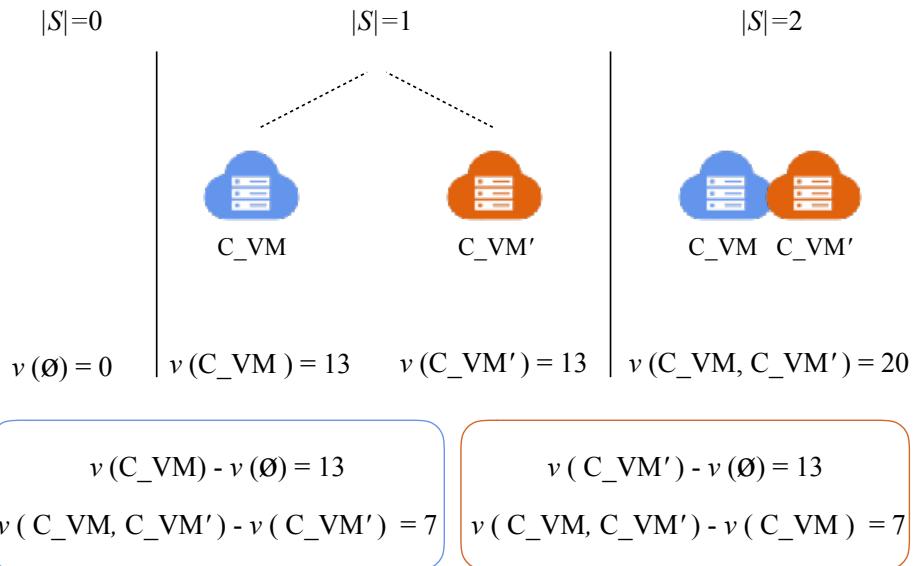


图 2-6 不同虚拟机集合体的能耗及边际贡献

为了更好地理解夏普利值方法的含义，本章仍旧使用图2-3中的虚拟机作为例子进行分析。显然，例子中虚拟机集合 $\mathcal{N} = \{C_VM, C_VM'\}$, $|\mathcal{N}| = 2$ 。图2-6展示了不同虚拟机集合体的能耗以及 C_VM 和 C_VM' 对这些集合体的能耗边际贡献。根据夏普利值的计算方法可得， C_VM 和 C_VM' 最终分配的能耗分别为：

$$\Phi_{C_VM} = \frac{13}{(2-0)*\binom{2}{0}} + \frac{7}{(2-1)*\binom{2}{1}} = 10 W,$$

$$\Phi_{C_VM'} = \frac{13}{(2-0) * \binom{2}{0}} + \frac{7}{(2-1) * \binom{2}{1}} = 10 W.$$

显然，夏普利值方法是一种基于边际贡献的分配策略，且在计算一个虚拟机能耗所得时，遍历了该虚拟机对所有集合体可能的边际贡献，并根据虚拟机集合体的基数不同，对不同的边际贡献分别进行不同的权值的加权。和表2.3对比可以发现，夏普利值方法的分配结果和理想的公平策略一致。

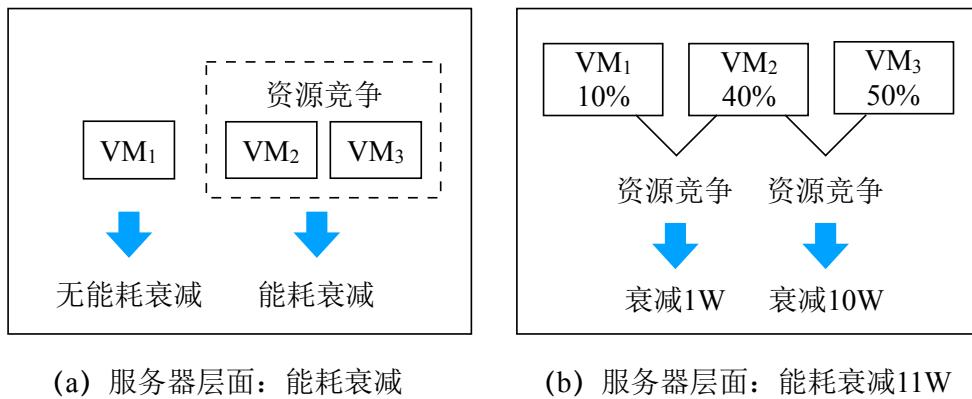


图 2-7 按资源使用比例进行能耗分配的公平性分析

此外，在上述例子中，按照虚拟机资源使用的比例进行能耗分配，也可以得到和夏普利值方法同样的结果，即 C_{VM} 和 $C_{VM'}$ 的 CPU 核心数量相同且利用率均为 100%，对总能耗 20W 按 1:1 比例进行分配，则各得到 10W 能耗。事实上，这种分配策略和夏普利值方法有着本质的不同，而且并不公平。图2-7展示了两种虚拟机竞争的场景。在场景 (a) 中， VM_2 和 VM_3 产生资源竞争导致能耗的衰减，最终形成反映在服务器层面的能耗衰减，而 VM_1 没有和另外两个虚拟机产生竞争，并没有使服务器的能耗有所减少。但是按照资源使用比例进行能耗分配时， VM_1 也会受益于 VM_2 和 VM_3 导致的能耗衰减。这显然对 VM_2 和 VM_3 并不公平。而在场景 (b) 中，由于资源竞争程度不同， VM_1 和 VM_2 导致能耗衰减 1w。然而按照虚拟机资源使用比例进行分配时， VM_1 的能耗实际被分配了 1.1W 的能耗衰减，这超过了 VM_1 实际产生的能耗衰减。和场景 (a) 类似， VM_1 也会受益于 VM_2 和 VM_3 导致的能耗衰减。因此，按照资源使用的比例进行能耗分配，并不是一种公平的方法。不同的是，夏普利

值方法是基于虚拟机加入一个集合体时，所带来的能耗边际贡献进行分配的。所以在场景（a）中，夏普利值方法可以捕捉到 VM_1 加入集合体 (VM_2 和 VM_3) 时，并没有产生任何能耗衰减的信息，同时在计算时也不会把 VM_2 和 VM_3 产生的能耗衰减分配给 VM_1 。类似地，在场景（b）中，夏普利方法也能捕捉到 VM_1 和 VM_2 导致能耗衰减 1W 的信息，并公平地进行能耗分配。因此，夏普利值方法是一种更加公平的方法。

2.3.3 四条公理在虚拟机能耗计量中的意义

夏普利值方法的公平性由于其满足的四条关于公平性的公理所决定的，而且这四条公理在虚拟机能耗计量中，有着重要的实践意义。

效率公理要求每个虚拟机估测的能耗值之和必须等于服务器的测量能耗。这条公理正好符合虚拟机能耗计量的宏观准确性条件。这对虚拟机能耗计价策略有着重要意义，保证了能耗计价策略不会导致运营商损失，也不会导致对用户的过度收费。显然，资源-能耗映射模型并不满足该公理，从图2-3可以看出，该模型可能导致对用户的过度收费，因此其并不适合用于虚拟机的能耗计量。

对称公理保证了两个处于不同物理机的虚拟机，在互相置换后且不会改变各自物理机能耗时，分配到相同的能耗相同，也就是虚拟机在能耗上是对称的。由于资源竞争导致虚拟机能耗的不确定性，就会存在资源利用率不同的两个虚拟机，实际对物理机的能耗贡献是相同的场景。由于资源-能耗映射模型是以资源利用率为标准进行能耗分配，最终会导致虚拟机因为资源利用不同，所分配的能耗不同，在上述场景中就会不满足能耗分配的对称性。

无效性公理是指如果一个虚拟机加入服务器并不改变服务器的能耗，则该虚拟机的能耗为零。通过实际测量发现，当虚拟机处于空闲状态时，并不会给服务器增加能耗。这是由于当虚拟机本身就是一种“软件应用”，在其空闲时几乎不对物理资源的进行实际操作，因而不会给服务器能耗造成影响。该公理保证了虚拟机在空闲状态下，其能耗为零。可以看出，无效性公理和对称公理都是“利益”分配过程，保证公平性的必要条件。

满足可加性公理使得能耗分配策略在虚拟机在使用跨服务器的资源时，也能够

同样适用。实际中，虚拟机可能会配置一个运行在其他服务器（如，磁盘阵列服务器）上的远程磁盘来扩展其存储空间，并可能和其他虚拟机共享同一个磁盘阵列服务器。这就使得虚拟机出现了跨服务器使用资源的情况。而满足该公理，可以分别计算同一个虚拟机在不同服务器上的能耗，其能耗总和即为虚拟机的总能耗，这使得一个虚拟机能耗分配策略保持局部结果之和和全局结果的一致性。

2.4 动态夏普利值方法

夏普利值方法曾被用于计量手机应用的能耗^[71]，然而在手机应用能耗计量中，由于手机应用是为了某一种功能而开发，计算资源使用模式比较单一，因此可以被当做是静态个体。与之不同的是，虚拟机提供的是一种通用处理能力，计算资源的使用模式变化大，而且对能耗的影响非常明显。因此在虚拟机能耗计量必须考量对应的状态。具体而言，本章提出了一种基于服务器的线性能耗模型，来克服虚拟机状态变化对夏普利值方法需要测量大量输入需求。

2.4.1 价值函数的状态扩展

尽管夏普利值方法在理论上非常适合虚拟机能耗计量，但是在实际应用并不容易。虚拟机的能耗高度依赖于虚拟机对服务器资源的使用情况，而且会随着时间不断发生变化。而虚拟机能耗计量的时间粒度一般要求秒级，因此虚拟机的能耗必须不断重复计算。对应地，夏普利值中的价值函数 $v(\mathcal{S})$ 还应该包含出集合体 \mathcal{S} 外的其他自变量，即集合体 \mathcal{S} 中各个虚拟机的状态。对此，本章提出一种新的动态夏普利值方法。

假设在任意一个时刻，一台服务器上有 n 个正在运行的虚拟机。虚拟机 $i, i \in \mathcal{N} = \{1, 2, \dots, n\}$ 则可以用其各个部件（如 CPU, 内存等）的状态向量

$$\mathbf{c}_i := [c_i^1, c_i^2, \dots, c_i^k] \quad (2.5)$$

来表示， k 代表了虚拟机可以获取到状态的部件数量。

对于任意一个虚拟机集合体 $\mathcal{S} \subseteq \mathcal{N}$ ，其价值函数 $v(\mathcal{S})$ 通过引入虚拟机的状态向量扩展为 $v(\mathcal{S}, \mathcal{C})$ ，其中 $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{|\mathcal{S}|}\}$ ，即

$$v(\mathcal{S}, \mathcal{C}) := v(\mathcal{S}, \mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{|\mathcal{S}|})。 \quad (2.6)$$

定义 2.1：对于一个服务器上运行的虚拟机集合 \mathcal{N} , 其对应的虚拟机状态为 $\mathcal{C}' = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{|\mathcal{N}|}\}$, 其能耗为 $v(\mathcal{N}, \mathcal{C}')$, 动态夏普利值方法通过引入虚拟机的状态向量, 将服务器能耗 $v(\mathcal{N}, \mathcal{C}')$ 公平地分配每个虚拟机的能耗 $\Phi_i(\mathcal{C}')$ 。

对应地, 动态夏普利值方法计算公式如下:

$$\Phi_i(\mathcal{C}') = \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \{i\}, \mathcal{C} \subseteq \mathcal{C}' \setminus \{\mathbf{c}_i\}} \frac{|\mathcal{S}|!(|\mathcal{N}| - |\mathcal{S}| - 1)!}{|\mathcal{N}|!} [v(\mathcal{S} \cup \{i\}, \mathcal{C} \cup \{\mathbf{c}_i\}) - v(\mathcal{S}, \mathcal{C})]。 \quad (2.7)$$

2.4.2 动态夏普利值的复杂度分析

一个虚拟机集合 \mathcal{N} 有 2^n 个子集, 因此动态夏普利值的计算存在两个方面的难题: (1) 夏普利值的计算复杂度为 $O(2^N)$, 因此其计算开销可能非常大; (2) 夏普利值方法的计算对应要求 2^n 个 $v(\mathcal{S}, \mathcal{C})$ 作为输入。

理论上, 只要服务器的资源足够, 其可以支撑任意数量的虚拟机同时运行。但是实际中, 出于性能的考虑, 一台服务器上运行的虚拟机数量往往是有限的。此外, 虚拟机的数量往往由服务器的 CPU 核心数量所决定。如在亚马逊弹性计算云中, 最低 CPU 配置的服务器为 16 核心, 其上运行的虚拟机最低配置为 1 核心, 而高配置的 32 核心服务器, 其上运行的虚拟机最低配置为 2 核心。因此在实际的云数据中心内, 一台服务器上运行的数量往往是有限的, 如亚马逊弹性计算云中, 一台服务器最多同时运行 16 个虚拟机。所以, 动态夏普利值在计算真实环境中的虚拟机能耗时, 计算开销并不高, 大约为 $2^{16} = 65536$ 次计算。

尽管动态夏普利值应用于虚拟机能耗计量的计算开销不高, 但是测量并收集 2^n 个 $v(\mathcal{S}, \mathcal{C})$ 作为输入, 在实际中不可能完成。举例来说, 要计算 16 个虚拟机组成的集合中各个虚拟机的能耗时, 需要将 16 个虚拟机的 2^{16} 个子集分别在服务器中运行并测量, 以获得动态夏普利值方法所需的输入。此外, 如果考虑每个子集中虚拟机的状态, 测量工作将会变得更加复杂。

2.4.3 基于同构虚拟机集合体的线性评估方法

针对动态夏普利值计算中获取所需输入 $v(\mathcal{S}, \mathcal{C})$ 的难题，本章提出了一种基于同构虚拟机集合体的线性评估方法。事实上， $v(\mathcal{S}, \mathcal{C})$ 代表的是当虚拟机集合体 \mathcal{S} 在服务器上运行时服务器的能耗（不包含静态能耗）。因此，获取 $v(\mathcal{S}, \mathcal{C})$ 的难题就可以转换为当任意一个虚拟机集合体 \mathcal{S} 以任意状态 \mathcal{C} 运行在服务器上时，如何获取服务器的能耗问题。从图2-2可以看出，线性资源-能耗映射模型对服务器能耗的估测非常准确，因此可以对价值函数进行能耗建模，从而消除大量的测量工作。

同构虚拟机集合体：虚拟机的配置不同会导致资源-能耗映射模型建模时，各个虚拟机的资源状态在数值上难以统一标准化。在云数据中心里，虚拟机的配置并不是任意的，而是有固定的配置。例如，亚马逊弹性云中的计算优化（Compute Optimized）虚拟机中的 $c4$ 类型，其只提供了 5 种配置方案。因此，同一台服务器上的虚拟机，可以根据其配置进行分类，配置相同的虚拟机组成一个同构虚拟机集合体。

一台服务器上的 r 种类型配置的虚拟机分别组成 r 个同构虚拟机集合体，进而可以重新定义价值函数如下：

$$v(\mathcal{S}, \mathcal{C}) := v(\mathcal{S}, \mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{|\mathcal{S}|}) := v(\mathcal{S}, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r), \quad (2.8)$$

其中 $\mathbf{v}_j = \sum \mathbf{c}_i, (j = 1, 2, \dots, r)$ 代表了同构虚拟机集合体 j 中所有虚拟机的状态向量之和。

最终，测量 2^n 个虚拟机子集的价值函数 $v(\mathcal{S}, \mathcal{C})$ 简化成了对 r 状态向量的服务器进行资源-能耗映射的建模。

线性评估方法：对于虚拟机集合体 \mathcal{S} 中任意一个同构虚拟机集合体 j ，其线性资源-能耗映射向量表示为 $\mathbf{w}_j = (w_j^1, w_j^2, \dots, w_j^k)$ ，则价值函数定义为：

$$v(\mathcal{S}, \mathcal{C}) = \sum_{j=1}^r \mathbf{w}_j \mathbf{v}_j. \quad (2.9)$$

假设已经测得的 m 个虚拟机集合的能耗 $\mathcal{V} = \{v(\mathcal{S}_1, \mathcal{C}_1), v(\mathcal{S}_2, \mathcal{C}_2), \dots, v(\mathcal{S}_m, \mathcal{C}_m)\}$ ，其中每个集合包含 r 种类型的虚拟机，则可通过最小二乘法求解如下优化问题来获

得资源-能耗映射向量 \mathbf{w}_j :

$$\underset{v(\mathcal{S}, \mathcal{C}) \in \mathcal{V}, \mathbf{v}_j \in \mathcal{C}}{\text{minimize}} \sum_{j=1}^r \|v(\mathcal{S}, \mathcal{C}) - \sum_{j=1}^r \mathbf{w}_j \mathbf{v}_j\|, \quad (2.10)$$

那么对于任意的虚拟机集合体 \mathcal{S}_x 及任意对应状态向量 \mathcal{C}_x , 其价值函数为

$$v(\mathcal{S}_x, \mathcal{C}_x) = \sum_{j=1, \mathbf{v}_j \in \mathcal{C}_x}^r \mathbf{w}_j \mathbf{v}_j. \quad (2.11)$$

通过求解上述最优化问题和公式 (2.11), 只需测量一部分虚拟机集合体状态以及对应服务器能耗, 即可获得所有的动态夏普利值方法所需的输入。

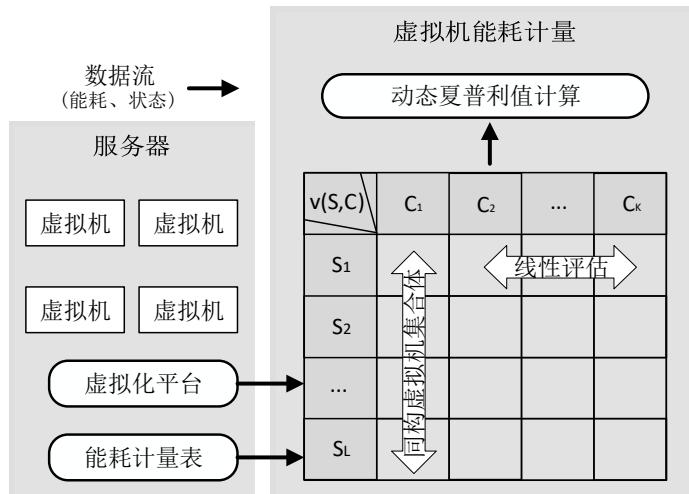


图 2-8 基于动态夏普利值方法的虚拟机能耗计量框架

2.5 性能测评

动态夏普利值方法是经典夏普利值方法的扩展, 其分配规则和经典的夏普利值方法是一致的, 因此理论上也是公平的。但是由于其输入 $v(\mathcal{S}, \mathcal{C})$ 没有办法完全通过真实测量获得, 因此本章提出的通过线性评估来获取输入的方法, 可能会导致结果与真实的夏普利值分配方法的结果有误差。本节通过部署真实的实验平台, 通过对服务器虚拟化, 对动态夏普利值方法进行测评。

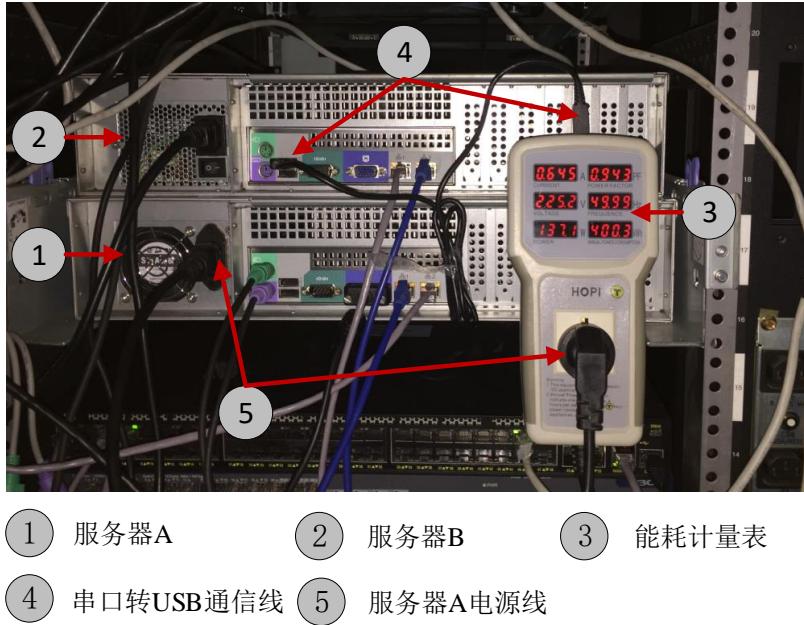


图 2-9 虚拟机能耗计量的实验平台

2.5.1 实验设置

图2-8展示了动态夏普利值方法在真实环境中部署的框架。该框架主要包含两个模块：离线数据收集模块和在线能耗评估模块。在离线数据收集模块中，虚拟化平台会收集到虚拟机运行过程中的部分状态，同时能耗计量表记录对应的服务器能耗信息，并将这两部分信息存储到 $v(\mathcal{S}, \mathcal{C})$ 表中。在线能耗评估阶段，虚拟化平台将当前虚拟机的状态推送给在线能耗评估模块。在线能耗评估模块通过查找 $v(\mathcal{S}, \mathcal{C})$ 表，并通过表中数据，根据2.4节中的动态夏普利值计算方法，对虚拟机的能耗进行评估。

图2-9展示了部署了上述框架的实验测量平台。该平台包含两台服务器（A 和 B）以及一个能耗计量表。服务器的配置均为 Intel Xeon 16 核心 CPU，32GB 内存，2TB 硬盘及 Linux 系统。服务器 A 运行 Citrix Xenserver 6.5 虚拟化平台，作为虚拟机运行的环境以及能耗评估对象，并连接能耗计量表。服务器 B 运行在线能耗评估模块，通过串口转 USB 通信线与能耗计量表连接，实时读取服务器 A 的能耗，同时收集服务器 A 中虚拟机的对应状态。能耗评估的时间间隔取决于虚拟机状态和能耗计量表的采样率，现有研究工作通常认为每秒评估一次称为实时评估^[28,32]。为保持一致，本章实验中采样率为 1 Hz。

表 2.4 虚拟机配置及对应的资源-能耗映射模型

虚拟机类型	CPU 核心数量	内存	硬盘	资源-能耗映射模型
VM ₁	1	2G	20G	$p = 13.15u$
VM ₂	2	4G	40G	$p = 22.53u$
VM ₃	4	8G	80G	$p = 50.26u$
VM ₄	8	14G	100G	$p = 96.99u$

实验中采用 *dstat* 工具进行虚拟机的状态收集。2.4节的动态夏普利值方法考虑了虚拟机的多种部件状态。根据实验平台的实际测量发现，CPU 的能耗变化在 0 到 160W 之间，而内存、磁盘的功耗在 10-12W 左右，且变化较小。因此，本实验主要采用 CPU 利用率作为虚拟机的状态。

实验的其他设置如下：

- 同构虚拟机集合体：根据服务器 A 的配置，实验采用了 4 种虚拟机配置。表2.4展示了 4 种虚拟机的配置以及对应测量得到的资源-能耗映射模型。对应地，可以将虚拟机分为 4 个同构虚拟机集合体。
- 测试程序：由于采用 CPU 利用率作为虚拟机的状态，实验选用了 SPEC CPU2006 测试工具中的若干 CPU 测试类型，作为验证动态夏普利值方法的测试程序。此外，实验中还采用了一种自制的程序，该程序会随机地使用 CPU，主要用于收集不同虚拟机 CPU 利用率下，服务器的对应功耗。表2.5展示了实验中具体使用的测试程序及其作用。
- 静态能耗：通过对实验平台的测试发现，当所有虚拟机处于关闭或者空闲状态时，服务器 A 的能耗相对稳定，保持在 137~139W 之间。因此，在实验测评中，服务器 A 的静态功耗以 138W 计算，并在和虚拟机能耗比较时，去除该部分静态能耗。

2.5.2 $v(\mathcal{S}, \mathcal{C})$ 的准确性评测

由线性评估所得的 $v(\mathcal{S}, \mathcal{C})$ 直接决定动态夏普利值的准确性，因此需要对输入的 $v(\mathcal{S}, \mathcal{C})$ 进行准确性评测。

表 2.5 实验测评中使用的测试程序

测试程序		功能描述	目的
整数测试	gcc (GC)	C 语言编译测试	验证实验结果
	gobmk (GO)	人工智能：围棋	
	sjeng (SJ)	人工智能：国际象棋	
	omnetpp (OM)	离散事件仿真	
浮点测试	namd (NA)	生物分子并行计算程序	
	wrf (WR)	天气分析预测	
	tonto (TO)	量子化学计算	
自制程序		随机地使用 CPU	测量不同的 $v(\mathcal{S}, \mathcal{C})$

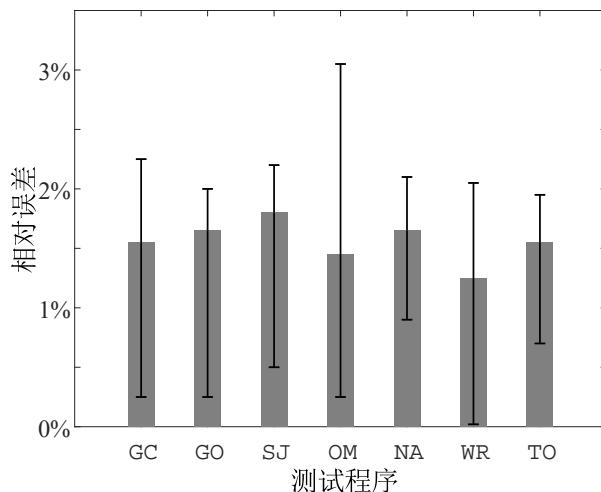


图 2-10 同构集合线性评估的误差

实验选择了两类虚拟机集合体。一种是由 4 个相同的 VM_1 类型虚拟机组成的集合，称为同构集合。另外一种是异构集合，分别由 1 个 VM_1 , 1 个 VM_2 , 1 个 VM_3 和 1 个 VM_4 组成。首先在服务器中运行 4 个同构的虚拟机并分别运行自制程序，并收集对应的状态和能耗数据。根据线性评估方法可得，其状态向量 $w_1 = 9.42$ 。类似地可以测得，四个异构的虚拟机的状态向量为 $[w_1, w_2, w_3, w_4] = [16.98, 17.91, 23.42, 75.21]$ 。

其次实验分别在同构集合和异构集合中运行 SPEC CPU2006 中的测试程序，并收集虚拟机的状态和服务器的能耗。通过线性评估得到的映射向量，将虚拟机的状

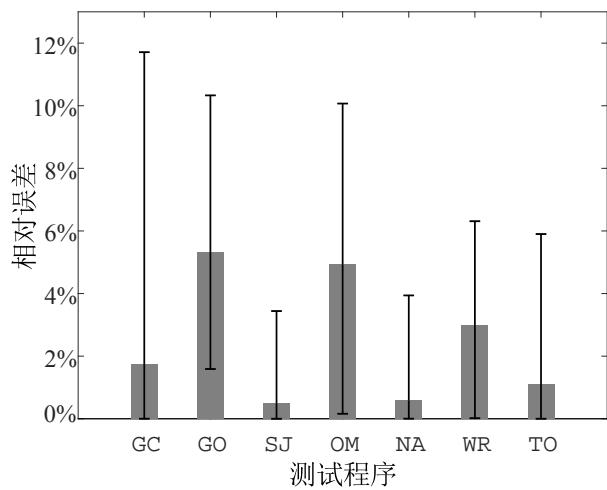


图 2-11 异构集合线性评估的误差

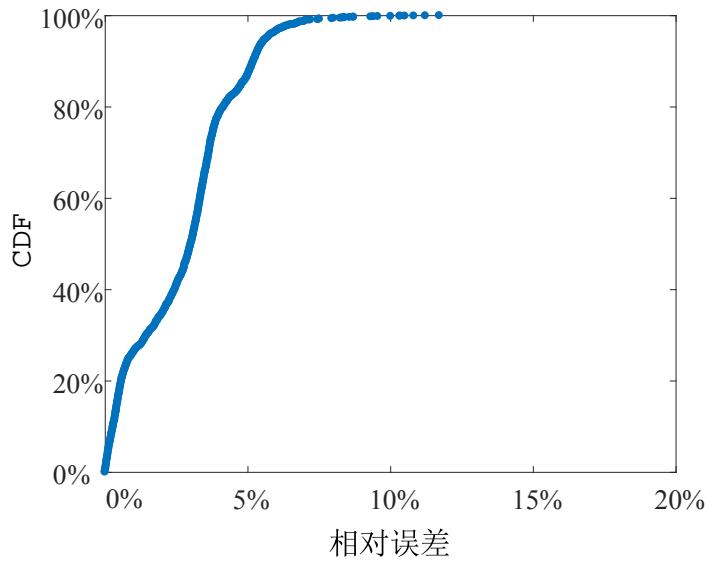


图 2-12 误差累积分布图

态转换为服务器的能耗，并和测量所得的服务器能耗作对比。图2-10和图2-11展示两者在同构和异构集中，运行不同测试程序时的相对误差。其中同构集合的平均误差为 1.79%，最大误差为 3.42%，异构集合的平均误差为 5.86%，最大误差为 11.7%。

为了更好地分析线性评估方法的准确性，实验进一步分析了相对误差的分布。图2-12展示了相对误差的累积分布图。在同构集合和异构集合两种情况中，超过 90% 的相对误差小于 5%，在不同测试程序中的平均相对误差为 5.33%。这说明使用线性

评估方法来获取动态夏普利值的输入是有效的。

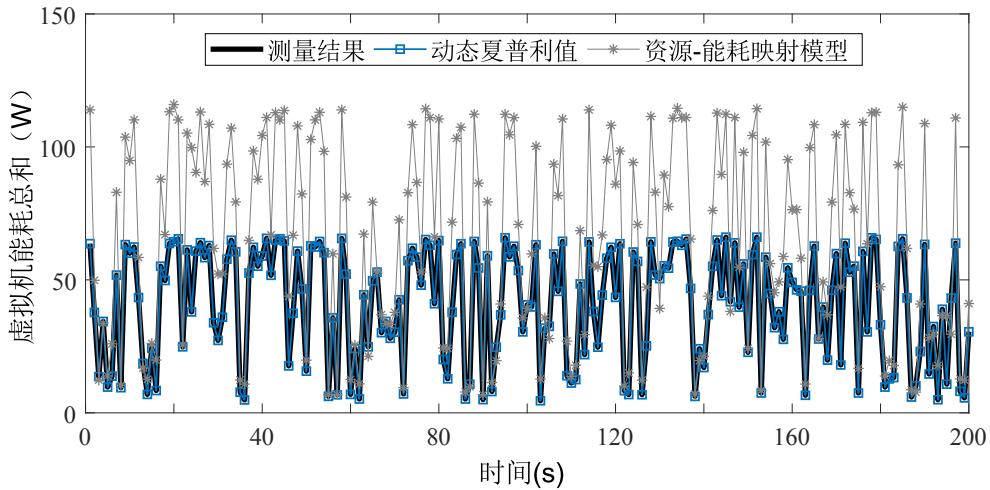


图 2-13 宏观准确性的比较

2.5.3 宏观准确性

此外，实验还比较了夏普利值法和资源-能耗映射模型的宏观准确性。在该实验中，服务器运行了 2 个 VM_1 类型，1 个 VM_2 类型，1 个 VM_3 类型和 1 个 VM_4 类型的虚拟机。图 2-13 比较了两种不同策略评估的虚拟机能耗总和与实际测量的服务器能耗（不包含静态能耗）。可以看到，资源能耗映射模型并不满足宏观准确性条件，其平均相对误差为 56.43%。值得注意的是，夏普利值方法由于满足效率公理，即使是线性评估方法得到的输入存在误差，其评估结果也始终与测量结果保持一致。由此可知，线性评估方法的误差并不会影响夏普利值方法的宏观准确性。这种误差可能对公平性造成一定影响。

2.5.4 虚拟机能耗计量中存在的其他挑战

尽管夏普利值方法为虚拟机能耗计量提供了一种有效手段，但是虚拟机的能耗计量仍然面临诸多挑战。

准确性问题：由于虚拟机本身是一种运行在物理硬件上的软件层面的逻辑抽象，当多个虚拟机共存于一台服务器时，虚拟机的真实能耗无法通过硬件（如能耗计量

表等) 进行测量, 只能通过软件方法进行估测。而正是由于缺少真实的能耗, 软件方法评估结果的准确性难以得到验证。因此, 目前还无法得到一种真正意义上准确的虚拟机能耗计量方法。

规模问题: 本章提出的方法只适用于一台服务器中虚拟机具有固定配置, 且数量有限的数据中心, 如亚马逊弹性计算云。而对于虚拟机配置是随机的(如, 用户自定义)的数据中心, 甚至在将来可能出现拥有超多核心 CPU(如, 128 核心)的服务器时, 夏普利值方法在处理大规模的虚拟机能耗计量问题时, 将会面临计算复杂度极高的挑战。

静态能耗分配问题: 静态能耗并不是有虚拟机直接产生, 而是服务器保持开启状态必然会产生。尽管如此, 静态能耗仍旧是为虚拟机正常的运行服务所产生的。如何在虚拟机之间公平地分配静态能耗, 仍旧时一个开放性问题。一般认为, 在虚拟机之间平均分配静态能耗, 或者按照虚拟机产生的动态能耗比例进行分配, 都是较为合理的方法。

2.6 本章小结

虚拟化技术提高了云数据中心的资源管理的便利性和有效性, 然而这也使得传统服务器层面的能耗监控无法满足虚拟化云数据中心的管理需求, 而虚拟机之间的资源竞争现象也使得传统的资源-能耗映射模型不再适用。另一方面, 虚拟机作为硬件资源的一种逻辑抽象, 无法直接对其进行能耗测量。因此, 很难找到一种可以直接验证的准确的能耗计量方法。针对此问题, 本章通过合作博弈理论, 将虚拟机能耗计量问题转化为经济学中的经典问题: 成本分配问题, 并提出了一种基于夏普利值的虚拟机能耗计量方法。该方法以宏观准确性和公平性作为虚拟机能耗计量方法有效性的评估条件, 为虚拟机能耗计量提供了一种新思路。此外, 为了适应虚拟机能耗的动态变化, 通过引入虚拟机的状态将经典的夏普利值扩展为动态夏普利值, 并利用线性评估方法, 有效地降低了夏普利值实际应用中的复杂度。

3 面向非 IT 设施层的能耗计量方法

上一章介绍了 IT 设施能耗如何在虚拟机之间进行公平划分。除了 IT 设施之外，非 IT 设施（如，制冷系统等）的能耗也占了云数据中心总能耗很大的比重。然而云数据中心只能监测到非 IT 设施的总体能耗，而且非 IT 设施的能耗增长和 IT 设施的能耗之间呈现一种非线性的关系，这使得现有的策略（如，平均分配，按 IT 能耗比例分配等）在进行非 IT 能耗分配时并不公平。针对此问题，本章提出了一种基于夏普利值的非 IT 能耗分配的方法。与上一章中不同的是，在云数据中心内一个非 IT 设施可能服务成百上千个虚拟机或服务器，而夏普利值方法的计算复杂度会随着虚拟机数量呈指数级增长。本章根据非 IT 设施的能耗增长特性，设计了一种轻量级的夏普利值能耗计量策略（Lightweight Energy Accounting Policy based on Shapley value, LEAPS），将夏普利值方法的复杂度从 $O(2^N)$ 降低到 $O(N)$ 。通过真实的数据中心能耗的测试验证，该方法与真实的夏普利值最大相对误差仅为 6.97%。

3.1 问题提出

云数据中心的 IT 设施依赖同样重要的非 IT 设施（如，UPS，制冷系统等）来保证服务器等设备的正常运行。在云数据中心里，这些非 IT 设施同样消耗大量的电力。在政府和绿色组织的压力下，越来越多的云用户开始关注自身的碳排放问题^[72]。举例来说，苹果公司（Apple）和阿卡迈（Akamai）等公司纷纷宣布要将自身云数据中心的能耗以及部署在第三方云数据中心业务所消耗的电力计入自身的碳排放中。因此，云数据中心的用户，不仅要计算 IT 设施的能耗，还要计算非 IT 设施的能耗。目前，工业界和学术界对 IT 设施的能耗计量提出了诸多方法，如服务器层面的能耗计量^[22]，虚拟机层面的能耗计量^[28,73]，软件应用层面的能耗计量^[74] 等。然而针对每个云用户所消耗的非 IT 设施能耗的研究，仍处于空白。

云用户一般以虚拟机的形式使用云数据中心中的计算资源，因此用户消耗的非 IT 设施的能耗，即为云数据中心中每个虚拟机所消耗的非 IT 设施能耗。云数据中

心中的非 IT 设施能耗主要由制冷系统和 UPS 所产生^[75]。根据全国数据中心信息开放平台的报告，2018 年我国大型和超大型数据中的 PUE 还分别高达 1.54 和 1.63^①。由于地理位置和气候等因素的限制，高能耗的制冷（如，制冷压缩机）方式仍旧是众多数据中心首选方案。这使得非 IT 设施的能耗占云数据中心总能耗的比例很大。根据数据中心制冷方案提供商的报告，即使是目前流行的水冷式系统，也只能减少 21%~22% 的制冷能耗^[76]。此外，目前电力系统中的 UPS 在进行电流转换时的效率也只有 80%~95%，导致云数据中心产生 5%~20% 的电力损失^[77,78]。在一个云数据中里，非 IT 设施的能耗在总能耗的平均占比为 30%~50%^[79,80]。

然而，计量每个虚拟机所消耗的非 IT 设施能耗是困难的。一个非 IT 设施由众多的虚拟机所共享，其能耗难以在物理层面或者逻辑层面进行有效的划分。显然，非 IT 能耗^②和 IT 能耗之间存在直接相关性，如服务器能耗增加会导致机房散热需求上升，从而导致制冷系统的能耗随之增加。一种经验主义的方法就是按照虚拟机 IT 能耗的比例来划分虚拟机的非 IT 能耗。但是这种能耗分配策略并不公平。因为非 IT 设施能耗与虚拟机的 IT 能耗之间呈非线性增长的关系，意味着即使是相同的 IT 能耗也会导致不同的非 IT 能耗。此外和服务器一样，非 IT 设施还具有静态能耗^[78,81]。如何公平地将这部分静态能耗在虚拟机之间进行分配也是一大难题。经验主义的方法虽然简单直观，但是缺乏理论依据，难以验证其结果的有效性。

本章针对非 IT 设施共享特性所导致的能耗划分难题，使用博弈理论将该问题建模成为成本分配问题^[82]。同第二章，该问题可以利用夏普利值方法进行求解，其中非 IT 设施的能耗即为“成本”，每个虚拟机为成本分配的对象。不同的是，在非 IT 能耗划分问题中一个非 IT 设施可能服务成千上万的虚拟机，因此参与分配的虚拟机的数量庞大，而夏普利值的计算复杂度为 $O(2^N)$ ，直接应用到该场景中会产生无法计算的难题。另一方面，通过测量发现，云数据中心内不同非 IT 设施的能耗与 IT 能耗之间的关系主要可以采用三种函数进行表达：线性函数，二次函数和三次函数。不同非 IT 设施的能耗特性不同，如何基于夏普利值建立统一的能耗划分方法也是一大挑战。针对以上问题，本章设计了一种轻量级的夏普利值能耗计量策略。具体来说，利

① 全国数据中心 PUE 情况，<http://www.odcc.org.cn/idccchina>

② 本文将虚拟机、服务器等 IT 设施产生的能耗称为 IT 能耗，对应地，非 IT 设施产生的能耗称为非 IT 能耗。

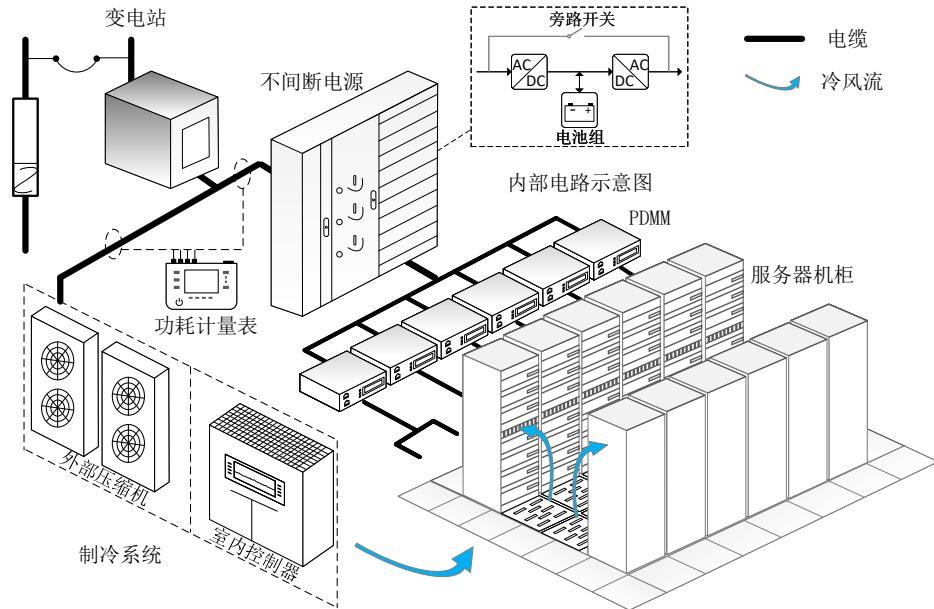


图 3-1 测量平台结构组成

用二次函数对不同的非 IT 设施能耗进行拟合，并通过推导证明得到一种和夏普利值等价的，复杂度仅为 $O(N)$ 的能耗计量方法。

3.2 非 IT 设施的能耗特性

本节主要通过测量和调研，介绍数据中心内非 IT 设施的能耗特性。

3.2.1 测量平台

本节测量了一个真实数据中心内非 IT 设施的能耗特征。测量平台为一个 346 个计算节点，16 个磁盘阵列，10 个 GPU 节点构成的小型数据中心，其尖峰功耗约为 166 kW。图3-1展示了该数据中心的整体构造。数据中心首先通过变电站，将电网的电力输送到 UPS 和制冷系统。UPS 首先将来自电网的高压交流电转换为低压直流电，以便对电池组进行充放电作为备用电源，然后再将电流转换为低压交流电给服务器进行电力供给^[83]。此外，该数据中心还配备了 7 个电力分配管理模块（power distribution management modules, PDMMs）。通过这 7 个电力分配管理模块可以获取到服务器的实时功耗（即 UPS 输出功耗）。由于位于城市中，该数据中心采用了精密空调制冷

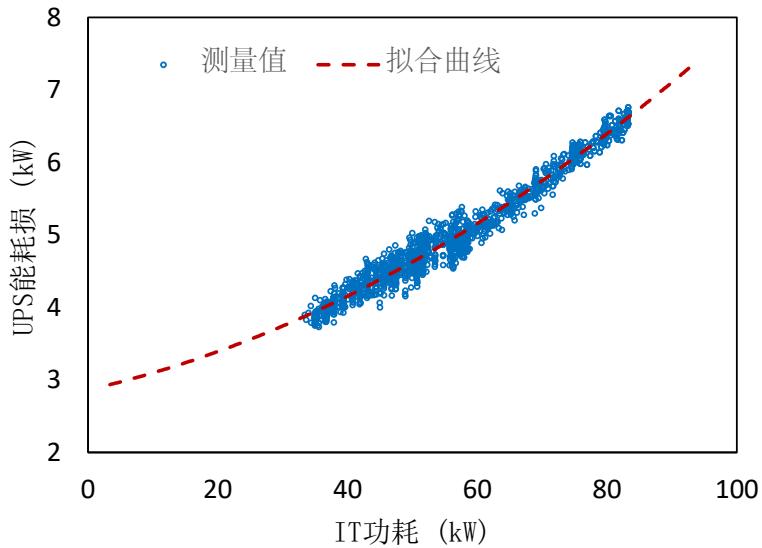


图 3-2 UPS 能耗损和 IT 能耗之间关系

来保障机房的温度环境。实验采用福禄克 (Fluke) 1738 三相功耗计量器来分别测量 UPS 的输入功耗和制冷系统的功耗。

3.2.2 UPS 能耗特征

由于需要对电流进行高低压和交直流的转换, UPS 会产生 5%~10% 的能耗损^[77,78]。根据施耐德对大型 UPS 系统的测量报告显示, UPS 的能耗损失增长和 IT 能耗的平方成正比^[78]。对应地, 实验对该数据中心的 UPS 系统进行了能耗测量, 其中能耗损可以通过 UPS 的输入功耗与输出功耗计算得到。

图3-2展示了 UPS 能耗损的测量结果。通过最小二乘法拟合可得, 能耗损和 IT 能耗之间的关系近似表达为:

$$F(x) = 0.0003x^2 + 0.0205x + 2.8628, \quad (3.1)$$

其中 x 代表了 IT 能耗, $F(x)$ 为 UPS 能耗损。以上的测量结果和施耐德公布的测试报告基本一致^[78], 该现象主要是由于 UPS 的电路发热产生的能耗损失和电流的平方成正比, 其中常数项为 UPS 空闲时消耗的电力^[80]。

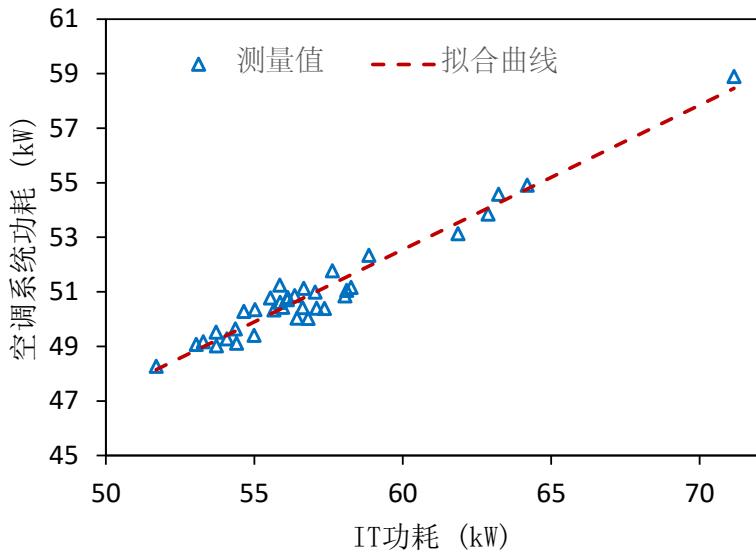


图 3-3 空调制冷系统能耗和 IT 能耗之间关系

3.2.3 制冷系统能耗特征

目前数据中心采用的制冷方式主要包括空调制冷，水冷以及自然冷却三种方式。

实验通过 45 天的测量，收集了在气温 30 摄氏度时，不同 IT 能耗对应的空调系统能耗。图3-3展示了测量结果，其中空调系统的能耗和 IT 能耗之间的关系可以近似表达为

$$F(x) = 0.53x + 20.749, \quad (3.2)$$

如图3-3所示，数据中心的能耗基本维持在一定范围内，因此测量范围仅在 50kW~80kW 之间。可以看出，空调系统的能耗和 IT 能耗之间呈线性关系。这主要是由于在外界温度固定的情况下，空调制冷系统的制冷量（即空调系统消耗 1 瓦特所能移除的热量）是固定的。由于服务器消耗的电能最终以热能形式散发到机房内，因此空调制冷系统的能耗和 IT 能耗之间呈线性关系。

水冷（Water Cooling）采用水等冷却液对服务器进行降温。冷却液通过循环管道输入服务器内，吸收 CPU 等部件的热量之后，再输送到服务器外进行冷却。根据研究发现^[80]，水冷系统的能耗和 IT 能耗的平方呈正比，具体可表达为: $F(X) = 742.8X^2 + 1844.6X + 538.7$ 。

自然冷却（Outside Air Cooling, OAC）直接利用室外的空气对服务器进行冷却。这种冷却方式一般要求数据中心建设在气温常年较低的地区并且需要气体过滤系统来去除灰尘等固体杂质，其能耗和室外气温有直接关系。研究发现^[79]，自然冷却方式的能耗可以表达为 $F(X) = kX^3$ ，其中 k 和室外气温相关。

本节通过真实测量和调研发现，数据中心中不同的非 IT 设施都和 IT 能耗之间存在直接相关性，且能耗特性和 IT 能耗之间的关系分别可以采用线性函数、二次函数以及三次函数进行表达^①。

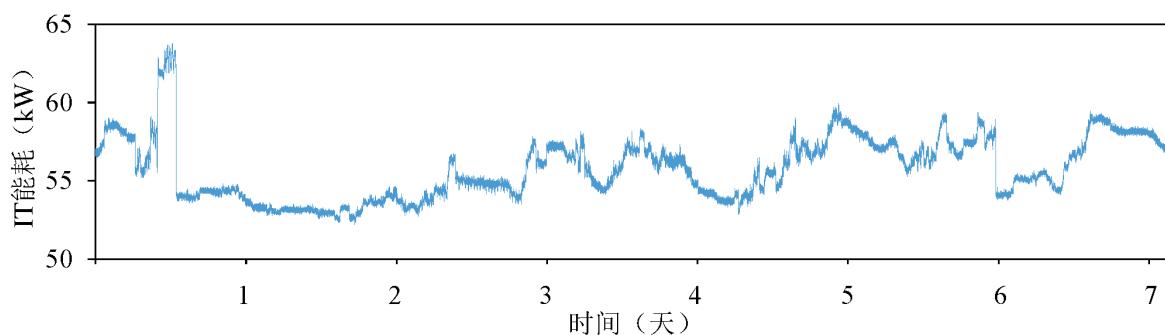


图 3-4 数据中心 IT 能耗负载变化

此外在测量过程中发现，数据中心在实际运行中 IT 能耗远远低于设计峰值。图 3-4 展示本节测量的数据中心一周内的 IT 能耗负载的变化，尽管该数据中心的设计峰值为 166 kW，但是从图中可以观察到该数据中心这一周内最高 IT 负载仅为 64 kW，这主要是由于服务器利用率低下造成的。这种现象使配套的非 IT 设施产生了极大的浪费，比如本节所测量的数据中心按照 IT 能耗的设计峰值配备了两个大型制冷压缩机，而在实际运行中却存在大量冗余的制冷能力没有被使用，产生了多余建设成本。然而并不能因此就减少非 IT 设施的设计容量，因为 IT 负载遇到突发峰值时，减少设计容量存在非 IT 设施超载的风险，严重的情况会导致服务器宕机甚至损坏。因此如何在应对不断变化的 IT 负载的同时，充分利用非 IT 设施冗余的能力来减少能耗，也是数据中心能效优化的可行途径。

^① 本章将上述三种函数统称为非 IT 设施的能耗特征函数。

表 3.1 非 IT 能耗计量问题中的主要变量及其含义

变量	含义
\mathcal{N}_j	非 IT 设施 j 所服务的虚拟机集合
$F_j(\cdot)$	非 IT 设施 j 的能耗特征函数
Φ_{ij}	非 IT 设施 j 为服务虚拟机 i 所产生的能耗
\mathcal{X}	集合 \mathcal{N}_j 的子集
P_i	虚拟机 i 的 IT 能耗
$P_{\mathcal{X}}$	子集 \mathcal{X} 中所有虚拟机的 IT 能耗之和
n_j	集合 \mathcal{N}_j 的基数
$r_{\mathcal{X}}$	子集 \mathcal{X} 的基数
δ_x	二次能耗特征函数在 IT 能耗为 x 时的误差
a_j, b_j, c_j	非 IT 设施 j 的二次能耗特征函数的系数

3.3 非 IT 能耗计量的定义

为了更好地描述非 IT 能耗计量中存在的问题与挑战，本节通过变量说明，具体定义了非 IT 能耗的计量问题。

3.3.1 问题定义

云数据中心通常以虚拟机的形式为用户提供计算服务^[84]，因此用户的非 IT 能耗计量可以转换为虚拟机的非 IT 能耗计量。假设在任意时刻，数据中心的一个非 IT 设施 j 同时服务一个虚拟机集合 \mathcal{N}_j 。根据上一节的测量结果，非 IT 设施 j 的能耗可以表示为 $F_j(\sum_{i \in \mathcal{N}_j} P_i)$ ，其中 P_i 代表了虚拟机 i 的 IT 能耗， $F_j(\cdot)$ 代表了非 IT 设施 j 的能耗特征函数。

定义 3.1： 非 IT 能耗计量的目标是寻找一种公平的方法，求出每个虚拟机所产生的非 IT 能耗 Φ_{ij} ，其中 $F_j(\sum_{i \in \mathcal{N}_j} P_i) = \sum_{i \in \mathcal{N}_j} \Phi_{ij}$ 。

非 IT 能耗计量问题主要面临两个方面的挑战。首先数据中心只能测量非 IT 设施的整体能耗，难以在物理层面和逻辑层面对其进行切割划分。此外非 IT 设施的能

耗增长是非线性的，如何设计一种统一有效的方法，同时保证能耗计量的公平性也是一大挑战。

3.4 夏普利值方法的优势与挑战

本节将介绍夏普利值方法如何应用于数据中心非 IT 能耗的计量问题及其优势，并与现有的三种经验主义的策略进行比较。

3.4.1 夏普利值在非 IT 能耗问题中的应用

同上一章类似，夏普利值方法同样可以应用到非 IT 能耗的细粒度计量中。特别地，本章中夏普利值方法中的价值函数为非 IT 设施的能耗特征函数 $F_j(P_{\mathcal{X}})$ ，其中 $P_{\mathcal{X}} = \sum_{k \in \mathcal{X}} P_k$ 。假设 $\mathcal{N}_j = \{1, 2, \dots, n\}$ 代表了 n 个由非 IT 设施 j 所服务的虚拟机集合，则一个虚拟机 i 从非 IT 设施 j 所分配到的非 IT 能耗为：

$$\Phi_{ij} = \sum_{\mathcal{X} \subseteq \mathcal{N}_j \setminus \{i\}} \frac{|\mathcal{X}|! (|\mathcal{N}_j| - |\mathcal{X}| - 1)!}{|\mathcal{N}_j|!} \cdot [F_j(P_{\mathcal{X}} + P_i) - F_j(P_{\mathcal{X}})] \quad (3.3)$$

3.4.2 公平性公理

上一章介绍了夏普利值的公平性是由其四条公平性公理所保证的，因此本章不再赘述。特别地，可加性公理在本章有新的含义：

(可加性公理 Additivity): 能耗的计量周期不能影响虚拟机的非 IT 能耗结果的一致性。意味着如果将一个非 IT 能耗计量周期 T 分割为多个能耗计量周期 $[t_1, t_2, \dots, t_n]$ ，那么虚拟机在每个时间间隔中所分配的非 IT 能耗之和，与在周期 T 的分配结果相同。即对于任意 $T = t_1 + t_2 + \dots + t_n$ ， $\sum_{t=t_1}^{t_n} \Phi_{i,j,t} = \Phi_{i,j,T}$ 。

3.4.3 现有策略的介绍及评价

策略一： $\Phi_{ij} = F_j / |\mathcal{N}_j|$: 该策略将非 IT 设施的总能耗，按照虚拟机的数量，进行平均分配，即每个虚拟机分配到的非 IT 能耗都相等。

策略二： $\Phi_{ij} = F_j \cdot P_{i \in \mathcal{N}_j} / \sum_{l \in \mathcal{N}_j} P_l$: 该策略将非 IT 设施的总能耗，按照虚拟机

表 3.2 三个虚拟机在不同时间内的 IT 能耗 (kW·s) .

虚拟机	t_1	t_2	t_3	$T = t_1 + t_2 + t_3$
#1	28	22	13	63
#2	24	9.2	10	43.2
#3	10	14.2	19	43.2

IT 能耗的比例进行分配。该策略常常应用在服务器托管数据中心^[63]。

策略三： $\Phi_{ij} = F_j(P_i + P_X) - F_j(P_X)$: 该策略将虚拟机加入时，对非 IT 设施能耗带来的边际增长作为虚拟机的非 IT 能耗。

以上三种策略虽然简单直观，但是事实上是不公平的。下面将从四条公平性公理出发，对以上三种策略进行评价。

策略一是将非 IT 能耗在虚拟机之间进行平分，而忽略了虚拟机之间的差别。每个虚拟机都会得到一个正数的非 IT 能耗，即便虚拟机没有实际的贡献，这违反无效性公理。

策略二违反了对称公理和可加性公理。以3.2节中的 UPS 为例，其能耗损和负载之间的关系可表示为 $F(x) = 0.0003 \cdot x^2 + 0.0205 \cdot x + 2.8628$ 。假设有三个虚拟机和一个能耗计量周期 $[t_1, t_2, t_3]$ (每个表示 1 秒)。则表3.2展示了虚拟机 #1，虚拟机 #2 和虚拟机 #3 不同时间内的 IT 能耗。根据 $F(x)$ 可以计算出在 $[t_1, t_2, t_3]$ 内总 UPS 能耗损为 $13.95\text{kW}\cdot\text{s}$ 。根据策略二，三个虚拟机按照各自 IT 能耗的比例，分配到的非 IT 能耗分别是 $5.84\text{kW}\cdot\text{s}$, $3.95\text{kW}\cdot\text{s}$ 和 $4.16\text{kW}\cdot\text{s}$ 。然而当能耗计量周期变换成 $T = t_1 + t_2 + t_3$ 时，表3.2中三个虚拟机的 IT 能耗分别为 $63\text{kW}\cdot\text{s}$, $43.2\text{kW}\cdot\text{s}$ 和 $43.2\text{kW}\cdot\text{s}$ 。根据策略二，这三个虚拟机分别分配到 $5.554\text{kW}\cdot\text{s}$, $4.03\text{kW}\cdot\text{s}$, $4.03\text{kW}\cdot\text{s}$ 。可以看到，策略二在使用不同的能耗计量周期时，会到的不一样的结果：对于虚拟机 #2， $\sum_{t=t_1}^{t_3} \Phi_t \neq \Phi_T$ ，这违反了可加性公理；而虚拟机 #2 和虚拟机 #3 在时间 T 中符合对称公理，但是在 $[t_1, t_2, t_3]$ 内分配的能耗不同，这违反了对称公理。可以看出，根据虚拟机的 IT 能耗比例来划分非 IT 能耗的策略本身是不能自洽的。

策略三违反了效率公理和对称公理。以两个虚拟机（表示为 #1 和 #2）为例，根据策略三，两个虚拟机的非 IT 能耗为别是 $\Phi_{1,j} = F_j(P_1 + P_2) - F_j(P_2)$ 和 $\Phi_{2,j} =$

$F_j(P_1 + P_2) - F_j(P_1)$ 。两者之和为 $\Phi_{1,j} + \Phi_{2,j} = 2F_j(P_1 + P_2) - F_j(P_1) - F_j(P_2)$, 如果满足效率公理, 则 $F_j(P_1 + P_2) = F_j(P_1) + F_j(P_2)$ 。显然, 当能耗特征函数 $F_j(\cdot)$ 是线性函数才能满足, 而根据3.2节, 能耗特征函数是非线性的。事实上, 策略三还有另外一种解释, 即虚拟机 #1 和 #2 是按顺序加入非 IT 设施 j 的。那么两个虚拟机的非 IT 能耗分别是 $\Phi_{1,j} = F_j(P_1) - F_j(0)$, $\Phi_{2,j} = F_j(P_1 + P_2) - F_j(P_1)$ 。这违反了对称公理, 例如当 $P_1 = P_2$, 则虚拟机 #1 和 #2 是对称的, 但是由于 $F_j(\cdot)$ 是非线性的, 因此 $\Phi_{1,j} \neq \Phi_{2,j}$ 。事实上, 在一个数据中心内存在成千上万的虚拟机, 很难区分虚拟机的加入顺序。策略三的第二种解释很难实现。因此在本章中, 策略三仅指第一种解释方法。

3.5 轻量级夏普利值能耗计量方法

夏普利值在非 IT 能耗计量问题的应用存在以下挑战: 夏普利值计算的时间复杂度为 $O(2^N)$, 即运算次数随着虚拟机的数量呈指数型增长。在一个数据中心内, 虚拟机的数量高达几千甚至几万, 这导致夏普利值无法计算。

3.5.1 基于二次函数的夏普利值化简证明

为了应对计算复杂度高的难题, 本章提出了一种基于夏普利值的轻量级能耗计量策略 (Lightweight Energy Accounting Policy based on Shapley value, LEAPS)。LEAPS 通过最小二乘法, 采用二次函数来拟合非 IT 设施的能耗特征, 即非 IT 设施 j 的能耗特征函数可以表达为:

$$F_j(x) = \begin{cases} 0, & \text{当 } x \leq 0 \\ a_j \cdot x^2 + b_j \cdot x + c_j, & \text{其他} \end{cases} \quad (3.4)$$

其中 x 代表了非 IT 设施 j 所服务的虚拟机集合的 IT 能耗, a_j , b_j 和 c_j 是通过最小二乘法拟合非 IT 设设施能耗特征, 所得的二次函数各项系数。

当虚拟机 i 的 IT 能耗为 0 时, 根据无效性公理和能耗特征函数, 该虚拟机对应的非 IT 能耗也为 0。因此本小节专注于讨论虚拟机 IT 能耗不为 0 的情况, 即非 IT 能

耗仅在 IT 能耗不为 0 的虚拟机之间进行分配。

当子集为空集时，即 $\mathcal{X} = \emptyset$ ，将拟合的二次函数公式 (3.4) 带入原始的夏普利值计算公式 (3.3) 中，并令 $|\mathcal{N}'_j| = n_j$ ，其中 $\mathcal{N}'_j \subseteq \mathcal{N}_j$ 代表了 IT 能耗不为 0 的虚拟机所组成的集合。则有

$$\begin{aligned} & \sum_{\mathcal{X}=\emptyset} \frac{|\mathcal{X}|!(|\mathcal{N}'_j| - |\mathcal{X}| - 1)!}{|\mathcal{N}'_j|!} \cdot [F_j(P_{\mathcal{X}} + P_i) - F_j(P_{\mathcal{X}})] \\ &= \frac{0!(n_j - 1)!}{n_j!} \cdot (a_j P_i^2 + b_j P_i + c_j - 0) \\ &= \frac{1}{n_j} \cdot (a_j P_i^2 + b_j P_i + c_j) \end{aligned} \quad (3.5)$$

当子集不为空集时，即 $\mathcal{X} \neq \emptyset$ ，将二次拟合函数 $F_j(x)$ 代入夏普利值原始计算公式 (3.3)，并令 $|\mathcal{X}| = r_{\mathcal{X}}$ 。则有

$$\begin{aligned} & \sum_{\mathcal{X} \neq \emptyset, \mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}} \frac{|\mathcal{X}|!(|\mathcal{N}'_j| - |\mathcal{X}| - 1)!}{|\mathcal{N}'_j|!} \cdot [F_j(P_{\mathcal{X}} + P_i) - F_j(P_{\mathcal{X}})] \\ &= \frac{2a_j P_i}{n_j!} \cdot \sum_{\mathcal{X} \neq \emptyset, \mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}} [r_{\mathcal{X}}!(n_j - r_{\mathcal{X}} - 1)!P_{\mathcal{X}}] \\ &+ (a_j P_i^2 + b_j P_i) \cdot \sum_{\mathcal{X} \neq \emptyset, \mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}} \frac{r_{\mathcal{X}}!(n_j - r_{\mathcal{X}} - 1)!}{n_j!}, \end{aligned} \quad (3.6)$$

其中 $P_{\mathcal{X}} = \sum_{k \in \mathcal{X}} P_k$ 。令 k 代表子集中的任意一个虚拟机，则所有的子集 $\mathcal{X} \setminus \{k\}$ 中，基数为 u 的非空子集共有 $\binom{n_j-2}{u-1}$ 个。因此，在所有的基数为 u 的子集 $\mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}$ 中，每个虚拟机 k 的 IT 能耗 P_k 出现了 $\binom{n_j-2}{u-1} = \frac{(n_j-2)!}{(u-1)!(n_j-u-1)!}$ 次，据此可得

$$\begin{aligned} & \sum_{\mathcal{X} \neq \emptyset, \mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}} [r_{\mathcal{X}}!(n_j - r_{\mathcal{X}} - 1)!P_{\mathcal{X}}] \\ &= \sum_{u=1}^{n_j-1} \sum_{\mathcal{X}, s.t., |\mathcal{X}|=u} [u!(n_j - u - 1)!P_{\mathcal{X}}] \end{aligned}$$

$$\begin{aligned}
 &= \sum_{u=1}^{n_j-1} \frac{(n_j-2)!}{(u-1)!(n_j-u-1)!} u!(n_j-u-1)! \sum_{k \in \mathcal{N}'_j \setminus \{i\}} P_k \\
 &= \sum_{u=1}^{n_j-1} u(n_j-2)! \sum_{k \in \mathcal{N}'_j \setminus \{i\}} P_k \\
 &= \frac{n_j!}{2} \sum_{k \in \mathcal{N}'_j \setminus \{i\}} P_k
 \end{aligned} \tag{3.7}$$

同理可知，在子集 $\mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}$ 中，基数为 u 的非空子集共存在 $\binom{n_j-1}{u} = \frac{(n_j-1)!}{u!(n_j-u-1)!}$ 个，因此

$$\begin{aligned}
 &\sum_{\mathcal{X} \neq \emptyset, \mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}} \frac{r_{\mathcal{X}}!(n_j - r_{\mathcal{X}} - 1)!}{n_j!} \\
 &= \sum_{u=1}^{n_j-1} \sum_{\mathcal{X}, s.t., |\mathcal{X}|=u} \frac{u!(n_j - u - 1)!}{n_j!} \\
 &= \sum_{u=1}^{n_j-1} \frac{(n_j-1)!}{u!(n_j-u-1)!} \frac{u!(n_j-u-1)!}{n_j!} \\
 &= \sum_{u=1}^{n_j-1} \frac{1}{n_j} \\
 &= \frac{n_j-1}{n_j}
 \end{aligned} \tag{3.8}$$

将公式 (3.7) 和公式 (3.8) 代入公式 (3.6)，再将公式 (3.5) 和 (3.6) 代入 (3.3)，最后得到与夏普利值等价的计算方法，即 LEAPS：

$$\Phi_{ij} = \begin{cases} 0, & \text{如果 } P_i = 0 \\ P_i \cdot [a_j \sum_{k \in \mathcal{N}'_j} P_k + b_j] + \frac{c_j}{n_j}, & \text{其他} \end{cases} \tag{3.9}$$

LEAPS 的最终推导形式给出了非 IT 能耗分配中的基本原则：静态非 IT 能耗 (c_j) 在虚拟机之间进行平分，动态非 IT 能耗按照虚拟机 IT 能耗比例进行划分（因为项 $a_j \sum_{k \in \mathcal{N}'_j} P_k + b_j$ 对于所有虚拟机相同）。显然，如果非 IT 设施的能耗特征函数符合

二次函数（如 UPS，水冷系统，空调制冷系统^①），则 LEAPS 和原始夏普利值方法是等价的。从上述推导可以看出，LEAPS 方法遵从了夏普利值方法的分配规则，区别是 LEAPS 采用了二次函数的拟合结果来作为夏普利值方法的输入。然而用二次函数去拟合具有三次函数能耗特征的非 IT 设施能耗（如自然冷却系统）是存在误差的。

3.5.2 误差分析

由于采用了二次函数拟合非 IT 设施能耗特征来作为夏普利值的输入，LEAPS 的计算结果可能会与真实的夏普利值方法产生误差。另一方面，由于夏普利值的计算复杂度非常高，其无法进行大规模计算。这使得 LEAPS 的结果失去了比较基准，其准确度难以有效验证。因此，本小节将具体分析 LEAPS 方法带来的误差及误差的量化方法。

假设当 IT 能耗为 x 时，真实的非 IT 能耗测量结果为 $F'_j(x)$ ，那么二次拟合函数与真实结果之间的误差可表示为 $\delta_x = F'_j(x) - F_j(x)$ 。因此， $F'_j(x)$ 可表示为

$$F'_j(x) = a_j \cdot x^2 + b_j \cdot x + c_j + \delta_x \quad (3.10)$$

将 $F'_j(x)$ 代入夏普利值计算公式中，根据上一小节的推导过程可得，真实的夏普利值可以表示为

$$\begin{aligned} \Phi'_{ij} &= P_i \cdot [a_j \sum_{k \in \mathcal{N}'_j} P_k + b_j] + \frac{c_j}{|\mathcal{N}'_j|} + \\ &\sum_{\mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}} \frac{|\mathcal{X}|! (|\mathcal{N}'_j| - |\mathcal{X}| - 1)!}{|\mathcal{N}'_j|!} \cdot (\delta_{P_{\mathcal{X}}+P_i} - \delta_{P_{\mathcal{X}}}) \end{aligned} \quad (3.11)$$

显然，LEAPS 和原始夏普利值方法的误差 (Δ) 为

$$\Delta = \sum_{\mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}} \frac{|\mathcal{X}|! (|\mathcal{N}'_j| - |\mathcal{X}| - 1)!}{|\mathcal{N}'_j|!} \cdot (\delta_{P_{\mathcal{X}}+P_i} - \delta_{P_{\mathcal{X}}}) \quad (3.12)$$

误差 Δ 的计算依旧具有很高的复杂度。但是从公式 (3.5) 和公式 (3.8) 可得

^① 线性函数可以归类为特殊的二次函数，其二次项系数 a_j 为 0。

$$\sum_{\mathcal{X} \subseteq \mathcal{N}'_j \setminus \{i\}} \frac{|\mathcal{X}|! (|\mathcal{N}'_j| - |\mathcal{X}| - 1)!}{|\mathcal{N}'_j|!} = 1 \quad (3.13)$$

该误差 Δ 可以解释为: $0 < |\mathcal{X}|! (|\mathcal{N}'_j| - |\mathcal{X}| - 1)! / |\mathcal{N}'_j|! < 1$ 是一个权重, Δ 是所有 $(\delta_{P_{\mathcal{X}}+P_i} - \delta_{P_{\mathcal{X}}})$ 的加权平均, 且所有权值之和为 1。

因此, 评估误差 Δ 可以转换为一个采样统计问题: 每个 $P_{\mathcal{X}}$ 是一个采样点, $(\delta_{P_{\mathcal{X}}}, \delta_{P_{\mathcal{X}}+P_i})$ 是一对采样值, 当采样规模为 $2^{|\mathcal{N}'_j|-1}$ 时, 采样所得 $(\delta_{P_{\mathcal{X}}+P_i} - \delta_{P_{\mathcal{X}}})$ 的加权平均是多少? 显然 $\delta_{P_{\mathcal{X}}}$ 和 $\delta_{P_{\mathcal{X}}+P_i}$ 决定了误差 Δ 的大小。下面将讨论并分析 $\delta_{P_{\mathcal{X}}}$ 和 $\delta_{P_{\mathcal{X}}+P_i}$ 产生的原因。

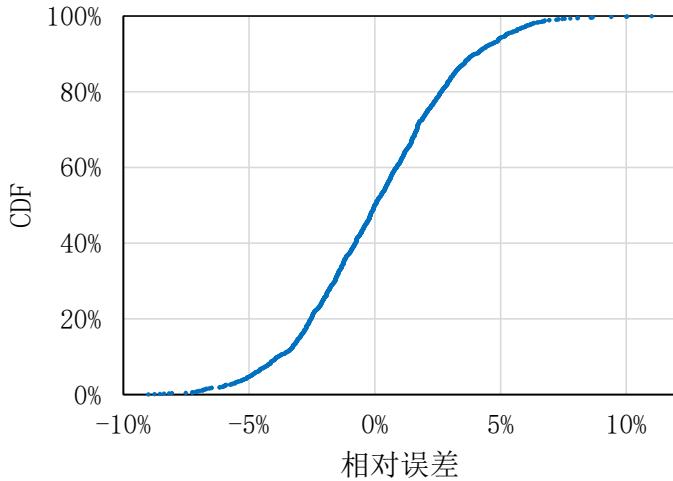


图 3-5 非确定性误差分布

对于符合二次函数能耗特征的非 IT 设施来说, $\delta_{P_{\mathcal{X}}}$ 和 $\delta_{P_{\mathcal{X}}+P_i}$ 主要由电气设施本身的波动产生。如图 3-2 所示, 尽管 UPS 的能耗特征符合二次函数, 但是测量所得的数据并不是完全落在拟合曲线上。由电气设施波动引发的误差具有随机性, 本章将这种误差称为**非确定性误差**。非确定误差虽然是随机的, 但是其符合一定的分布规律。图 3-5 展示了误差分析的结果, 非确定性误差近似地符合一个正态分布, 其中 $\mu = 0, \sigma = 0.023$, 而且约 95% 的误差小于 4.6%。由于非确定性误差本身较小, 因此可以合理推测由非确定性误差引起的 Δ 也较小。后续章节将会通过实验测试, 验证这种推测的正确性。

对于符合三次函数能耗特征的非 IT 设施来说, 用二次函数去拟合其能耗特征, 必

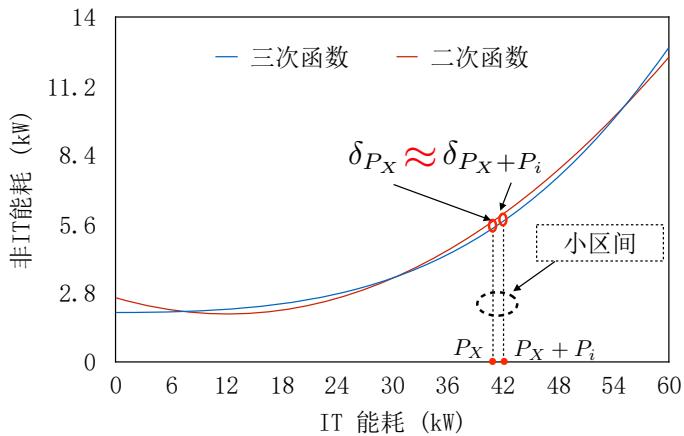


图 3-6 二次函数拟合三次函数

然会产生误差。这种误差称为**确定性误差**。图 3-6展示了使用二次函数 ($F = 0.0046x^2 - 0.1126x + 2.5957$) 拟合三次函数 ($F = 0.00005x^3 + 2$)，在区间 $0 < x < 60$ kW 的结果。确定性误差是指三次函数和二次函数之间的差值，即图 3-6中的 δ_{P_X} ， $\delta_{P_X+P_i}$ 。与一个数据中心的总 IT 能耗相比（几十千瓦甚至兆瓦），一个虚拟机的能耗很小（约为 0~0.3kW）。而 P_i 代表了一个虚拟机的能耗，因此 $[P_X, P_X + P_i]$ 是一个很小的区间。这意味着 $\delta_{P_X+P_i}$ 和 δ_{P_X} 的值非常接近，因此两者之差也趋近于零。因此，任意选取一个 P_X ， $(\delta_{P_X+P_i} - \delta_{P_X})$ 是一个趋近于零的值，而 Δ 是这些值得加权平均，因此 Δ 也是一个趋近于零的值。

根据以上分析，不管是非确定性误差还是确定性误差，由其产生的 LEAPS 和原始夏普利值之间的误差始终较小。为了进一步验证，实验将通过对非确定误差和确定性误差进行仿真，并通过统计分析验证 LEAPS 在实际应用到大规模非 IT 能耗计量时的准确度。

3.6 性能测评

3.6.1 实验设置

实验设置两种类型的非 IT 设施，其能耗特征分别可以用二次函数和三次函数进行表示，具体参数设置如下：

- 能耗计量周期：两次非 IT 能耗计量的间隔设置为 1 秒，保持 IT 能耗计量的时

间周期一致^[22]，也称为实时计量。

- IT 能耗: 实验测量了3.2节中数据中心的 IT 能耗，并用此真实的 IT 能耗进行实验评估。
- UPS 能耗特征: 实验采用了3.2节数据中心的 UPS，其能耗特征为二次函数，具体表达式为 $F(x) = 0.0003x^2 + 0.0205x + 2.8628$ 。
- OAC 能耗特征: 对于能耗特征为三次函数的非 IT 设施，实验采用了相关工作中测得的 OAC 系统能耗特征^[79]，当室外气温为 30 摄氏度时，其表达式为 $F(x) = 0.0005x^3$ ，对应地二次拟合函数为 $F(x) = 0.0457x^2 - 0.1255x + 5.9566, 0 < x < 60$ 。
- 非确定性误差: 3.5.2节分析了非确定性误差产生的原因，并且发现非确定性误差服从一个 $\mu = 0$ 且 $\sigma = 0.023$ 的正态分布。因此可以在拟合的二次函数基础上，加上随机产生的正态分布误差后，作为原始夏普利值的输入，仿真出真实环境下的夏普利值。LEAPS 通过拟合的二次函数计算出的结果和仿真结果进行比较。

表 3.3 LEAPS 和夏普利值计算时间比较

虚拟机数量	LEAPS 计算时间	夏普利值计算时间	夏普利值改进算法
10	0.009 ms	33.59 s	0.003 s
20	0.017 ms	> 1 天	5.57 s
10000	2.1 ms	无法计算	无法计算

3.6.2 计算复杂度比较

表3.3展示了在一台配置 Intel Xeon E5-2650 V4 CPU 的服务器上，LEAPS，夏普利值和夏普利值改进算法计算时间比较。夏普利值在虚拟机数量超过 20 时，其计算时间就超过一天，即使使用改进算法^①，也难以在实际的数据中心内应用。本章提出的 LEAPS 方法，在虚拟机数量高达 10000 时，其计算时间只需 2.1ms，能有效地对

^① Shapley value (fast), <https://ww2.mathworks.cn/matlabcentral/fileexchange/57735-shapley-value-fast>

虚拟机的非 IT 能耗进行实时计量。

3.6.3 LEAPS 方法准确度

从上一小节的计算复杂度可以看出，在虚拟机数量比较小时，原始夏普利值方法能在可忍受时间内计算出结果。因此 LEAPS 的准确度在极小规模下可以得到验证。但是当虚拟机数量较大时，原始夏普利值方法的结果无法得到，因此验证 LEAPS 在大规模虚拟机场景下的准确度存在困难。但是通过3.5.2节分析可知，LEAPS 的误差验证可以转换成一个采样统计的问题，即每个 P_x 是一个采样点， $(\delta_{P_x}, \delta_{P_x+P_i})$ 是一对采样值，当采样规模为 $2^{|\mathcal{N}_j|-1}$ 时，采样所得 $(\delta_{P_x+P_i} - \delta_{P_x})$ 的加权平均是多少？

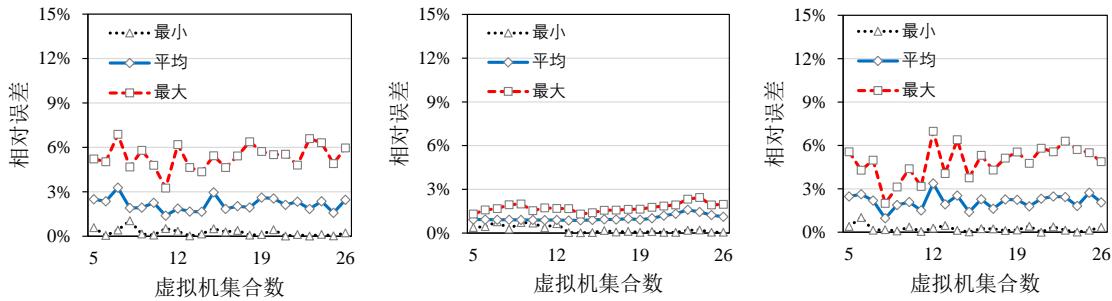


图 3-7 非确定性误差对 Δ 的影响 图 3-8 确定性误差对 Δ 的影响 图 3-9 非确定性 + 确定性误差对 Δ 的影响

据此，实验首先将虚拟机分成 5 个虚拟机集合，即将总 IT 能耗随机拆分为 5 份，将每个虚拟机集合作为非 IT 能耗的计量对象。对应地，可以仿真出夏普利值在虚拟机规模为 5 时的结果，和 LEAPS 结果进行比较。然后再将总 IT 能耗拆分为更多份，即增加虚拟机数量规模，再仿真夏普利值。图3-7-图3-9展示了虚拟机规模从 5 到 26 时，LEAPS 和夏普利值之间在不同虚拟机规模下的误差。尽管虚拟机的数量很少，但是 $(\delta_{P_x}, \delta_{P_x+P_i})$ 区间 $[0, 54.872]$ 上的采样数量从 32 增长到了超过 3350 万。当采样数量按照 $[32, 64, 128, \dots, 33554432]$ 变化时，LEAPS 的误差并没有增大，其平均误差（即， $\delta_{P_x+P_i} - \delta_{P_x}$ 的加权平均值）仅为 3.28% 和 3.37%，最大误差为 6.88% 和 6.97%。此外如图3-8所示，确定性误差对 LEAPS 的准确性影响很小，平均误差仅为 1.59%。以上实验结果验证了3.5.2节对 LEAPS 误差的分析是正确的。对于非确定性误差，由于其本身较小，因此加权平均后依然很小。对于确定性误差，由于 $(\delta_{P_x+P_i} - \delta_{P_x})$

的值趋近于 0，因此加权平均后的误差仍然很小。从实验结果可以看出，采样数量对 LEAPS 的准确度没有影响，因此即使是虚拟机数量高达成千上万时，LEAPS 的准确度依旧很高。

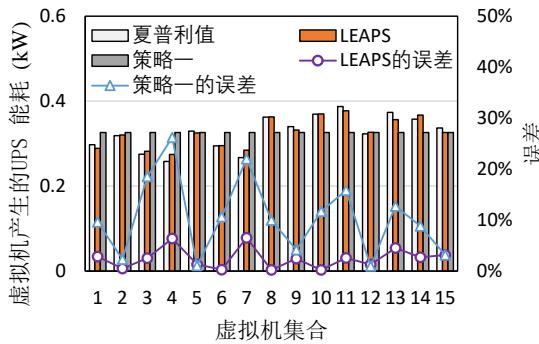


图 3-10 与策略一 UPS 能耗计量的对比

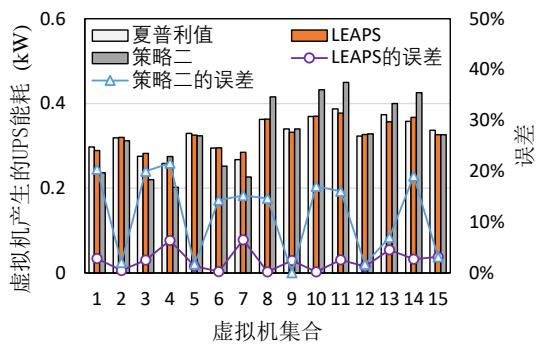


图 3-11 与和策略二 UPS 能耗计量的对比

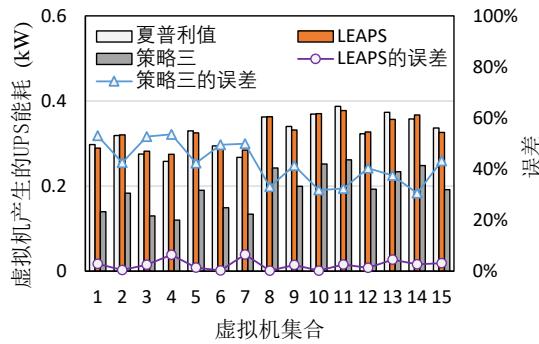


图 3-12 与策略三 UPS 能耗计量的对比

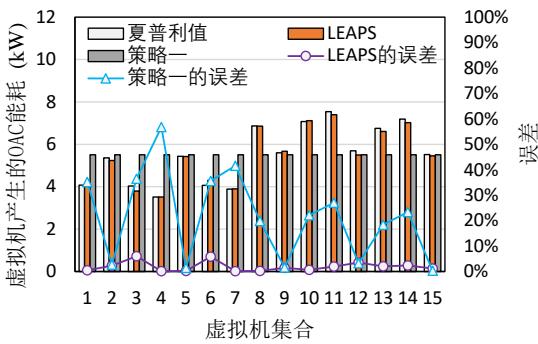


图 3-13 与策略一 OAC 能耗计量的对比

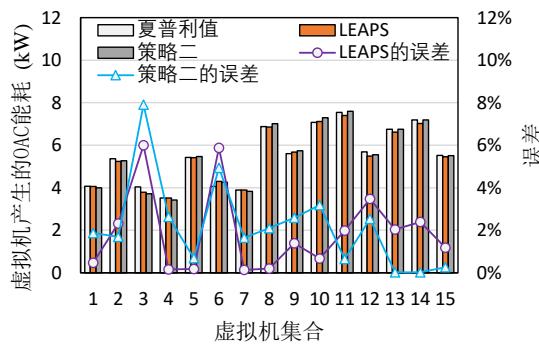


图 3-14 与策略二 OAC 能耗计量的对比

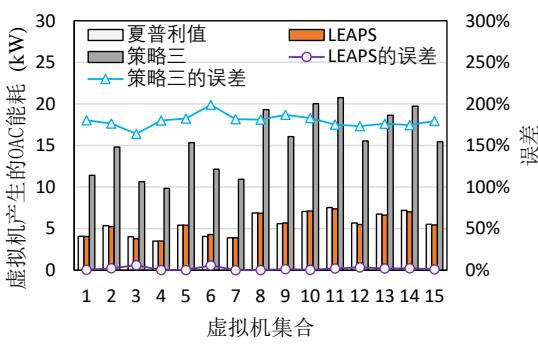


图 3-15 与策略三 OAC 能耗计量的对比

3.6.4 LEAPS 和其他策略的比较

由于夏普利值的计算复杂度极高，而且前文已经论证本章提出的方法的误差和并不会因虚拟机数量而变化，因此将虚拟机分成 15 个虚拟机集合，作为非 IT 能耗的统计对象，同时计算出不同策略（即3.4.3节中的策略一至策略三）的计量结果，和夏普利值方法的结果进行比较。图3-10至图3-15展示了不同策略和夏普利值的比较结果。可以看出，策略一至策略三的计量结果和夏普利值结果相比误差较大，而 LEAPS 计量 UPS 能耗的最大误差仅为 6.53%，计量 OAC 能耗的最大误差仅为 5.92%。

策略一将能耗在不同虚拟机之间进行平分显然是不公平的。而策略二和 LEAPS 方法最大的不同就是 LEAPS 将静态能耗在虚拟机之间进行平均分配。因此在对 OAC 进行能耗计量时，由于 OAC 没有静态能耗，所以策略二和 LEAPS 方法的结果相似。策略三的能耗计量基于边际增长，其在能耗计量的过程中忽略了静态能耗，因此其对 UPS 能耗的计量结果小于其他策略。

3.7 本章小结

非 IT 设施作为数据中心重要的组成部分，其能耗占比不容忽视。本章针对非 IT 能耗计量问题，本章提出了一种基于夏普利值的非 IT 能耗计量方法。此外针对夏普利值计算复杂度极高的问题，本章利用非 IT 设施的能耗特性，通过推导证明得到一种基于夏普利值方法，但是复杂度仅为 $O(N)$ 的非 IT 能耗计量方法。该方法最终的推导形式给出了具有理论依据，但是复杂度极低的能耗分配原则：静态非 IT 能耗在虚拟机之间进行平分，动态非 IT 能耗按照虚拟机 IT 能耗比例进行划分。

此外本章测量中发现，数据中心 IT 能耗在实际运行中远远低于设计峰值，导致配套非 IT 设施利用率和能效低下。本文将在后续章节中围绕如何充分利用非 IT 设施的冗余能力，对数据中心的能效优化进行进一步探究。

4 基于 IT 负载感知的混合水冷系统

上一章测量发现 IT 负载低下导致制冷系统存在大量的冗余制冷能力，而且 IT 负载不断发生变化。这种现象使得冗余的制冷能力只能用于应对突发峰值，而在其他时候存在大量浪费，对此本章结合水冷技术实现了一种 IT 负载和制冷动态控制相结合的能效优化系统，该系统充分利用水冷系统冗余的制冷能力所提供的调节空间，通过提高供水的温度（称为温水制冷）来降低数据中心制冷能耗。同时系统针对温水制冷导致冗余制冷能力大幅减少以及在服务器负载高峰存在制冷失败的风险（即水温过高无法有效冷却服务器），将半导体制冷片（Thermoelectric Cooler, TEC）整合到现有的水冷系统，提供细粒度的制冷。通过真实系统原型和数据中心数据集的验证，本章提出的混合水冷系统对于 CPU 的制冷效率，达到了 1.04~1.05 的局部 PUE。

4.1 问题提出

近年来，新型的水冷系统越来越受到数据中心运营商的关注。和传统风冷系统相比，水冷系统具有更高的制冷效率，不但可以提高数据中心内服务器的密度，更重要的是，水冷系统可以通过自然蒸发进行散热，可以有效降低制冷系统的能耗，提

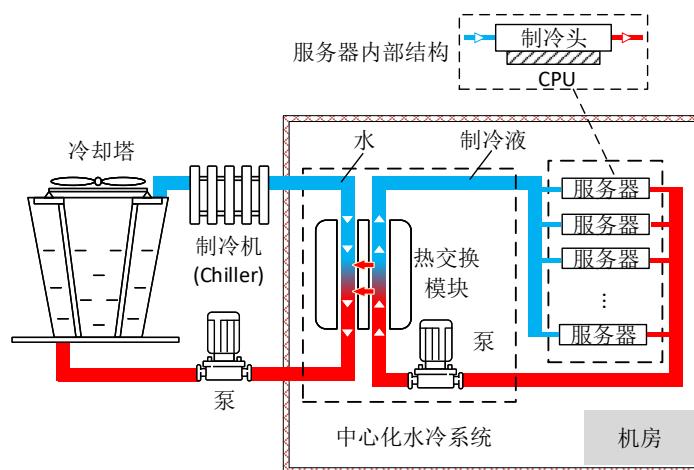


图 4-1 数据中心水冷系统结构

高数据中心能效^[85]。如图4-1所示，数据中心水冷系统结构主要包含两个循环：内部制冷液循环和外部水循环。内部循环的制冷液主要由去离子水构成，防止对管道造成腐蚀而泄露。制冷液流入每个服务器，在服务器内部通过制冷头吸收CPU的热量，再将热量带到热交换模块。外部循环一般采用普通自来水，通过热交换模块吸收制冷液的热量，再将热量带到冷却塔。冷却塔通过自然蒸发等手段将热量散到室外空气中。这个过程中，热量交换主要通过自然的热传导完成，几乎不消耗能量。水在冷却塔中散去一部分热量后，还需要经过制冷机，进一步降低温度。根据主流数据中心水冷方案提供商的报告显示^[76]，上述结构的水冷系统可以同时为数千台服务器提供制冷，并且和空调制冷相比，水冷系统可以节约21%至22%的制冷能耗。

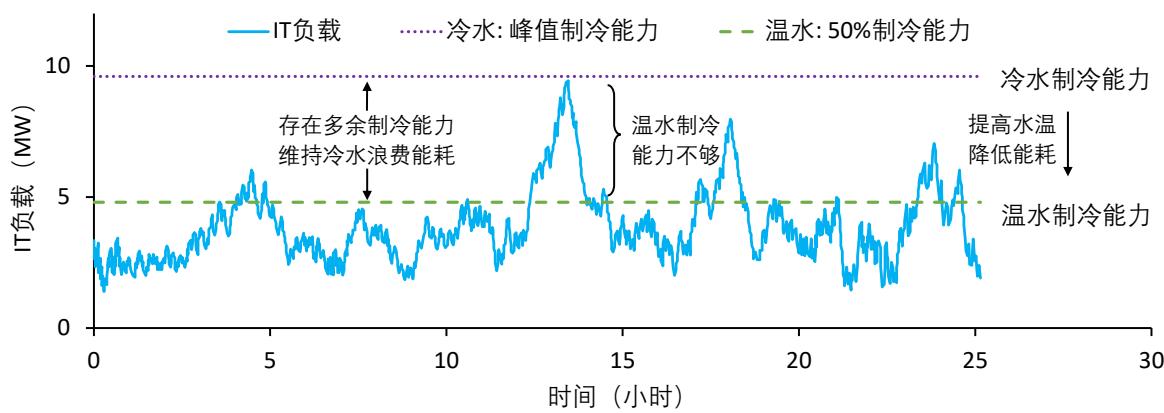


图 4-2 冷水和温水制冷策略的优缺点比较

当前数据中心的水冷采用的策略非常保守，一般采用7至10摄氏度的水来对服务器进行制冷^[86]。正如上一章测量发现数据中心的服务器利用率低下，这种现象在成熟的商业数据中心中也普遍存在，如阿里巴巴数据中心平均利用率不超过25%^[87]，谷歌数据中心平均利用率不超过20%^①，这使得水冷系统存在大量冗余的制冷能力。另一方面冷却塔散热能力有限，往往需要制冷机进一步降低水温，而制冷机和空调原理类似，是通过消耗电力进行制冷，这使得数据中心水冷系统的能耗大大增加，尤其是在高温天气和地区。因此充分利用冗余的制冷能力所提供的可调节空间，提高水温减少制冷机的使用（即温水制冷），使CPU的运行温度在安全范围内适当提高，

^① Google cluster workload traces, <https://github.com/google/cluster-data>

是一种有效的节能手段。据施耐德公司报告显示^[86]，将水温要求从 7 至 10 摄氏度提高到 18 至 20 摄氏度可以使水冷系统节能 49% 左右。图4-2利用谷歌公司数据中心的能耗数据^[48]，通过例子展示了采用温水和冷水优缺点。从图中可以看到，如果采用能够提供峰值制冷能力的冷水，会在大部分时候存在冗余的制冷能力，而维持冷水会浪费大量能耗。温水虽然减少了冗余制冷能力和能耗浪费，但是存在制冷失败的风险。虽然数据中心平均利用率较低，但是仍然会出现短时间内负载很高的现象。尤其当服务器负载突然升高时，CPU 的温度会在几秒内迅速上升，而水冷系统需要较长时间才能降低水温，并把水输送到对应地服务器，这使得温水制冷存在一定风险。

另一方面，出于工程复杂度和成本考虑，当前数据中心水冷系统采用的都是中心化控制（即通过控制全局水温来调控制冷能力）^[88]。由于无法提供细粒度的制冷能力，大大制约了水冷系统的能耗效率。在温水制冷中，存在一部分服务器超过了安全温度阈值，而另一部分没有超过安全值的局部热点现象。中心化控制的水冷系统必须按照局部热点来控制全局水温，而其他的非热点服务器可能并不需要额外制冷。这就会产生不必要的制冷，造成能耗浪费。显然，如果能解决温水制冷在面对突发负载导致温度过高的风险和局部热点制冷效率的问题，数据中心的制冷能耗开销将大大降低。

目前存在诸多通过任务负载调度来解决数据中心制冷难题的方案，称为温度感知的任务负载调度。数据中心的热量主要由服务器功耗的高低决定^[89]，而服务器的功耗和其任务负载相关。因此通过推迟任务的执行可以降低一定时间内的制冷需求。任务推迟可以一定程度降低制冷需求的峰值。但是上述方案不适用于对完成时间敏感的任务。交互型任务需要实时响应，任务推迟会降低服务质量。另外服务器在开机状态下，即使没有执行任务，也会产生“静态”能耗^[22]。因此通过任务负载整合，将任务集中到一部分服务器并关闭利用率低的服务器，可以减少服务器能耗，从而降低整体的制冷需求。但是任务整合后的服务器由于利用率升高，对水温的要求也提高，这导致制冷机使用更加频繁，因此服务器功耗优化并不意味着制冷功耗的优化^[90]。此外，任务整合并不适合交互性任务负载。由于排队效应，在服务器利用率提高后，任务的响应时间会变长^[60]。比如网页搜索等交互性服务要求不管流量多少，所有服务器都需要保持开机待命状态^[91]。动态 CPU 频率调控（Dynamic Voltage Frequency

Scaling, DVFS) 可以通过降低 CPU 频率, 来减少 CPU 的能耗和热量, 但是这种方法并不适合性能保证的任务。比如亚马逊弹性计算云中的计算优化型实例 *c5* 保证至少提供用户 3.0 GHz 的 CPU 频率^①。此外, 任务负载均衡被认为可以平衡服务器功耗和散热需求。但是任务负载的目标通常是提高性能(如缩短任务完成时间等), 其平衡通常代表任务处理速度的平衡, 并不是服务器利用率的平衡(如异构服务器环境中), 因此任务平衡并不一定能带来温度的平衡。目前服务器利用率不均衡在数据中心仍旧是一种常见现象^[92]。基于软件方法的数据中心制冷优化方案可能还需要提前预测任务的资源需求、完成时间等作为调度依据^[60,93], 最终效果非常依赖预测算法准确性。

在另一方面, 温度感知的任务负载调度可能会和数据中心已有的任务调度策略相冲突, 如网络感知的任务调度^[94,95], 需要在系统性能和制冷能效之间妥协平衡。而在真实的数据中心运维中, 任务调度和制冷系统通常由不同的工程师或者部门负责。负责任务调度的工程师更加关注于提高系统性能而非制冷系统能效, 导致温度感知的任务调度在实际中往往难以部署。

针对温水制冷中存在的挑战和任务负载调度存在的局限性, 本章提出一种混合液冷系统, 该系统可以实现安全的温水制冷, 并能随着任务负载的变化提供最佳能效的制冷策略。在硬件上, 该系统在原有水冷系统的基础上, 将半导体制冷片(TEC)整合到每个服务器中, 为每个服务器提供实时、细粒度的制冷。在软件上, 设计了自适应的制冷控制方法, 能随着任务负载变化, 使用制冷机和 TEC 提供混合制冷方案, 以达到最佳能效。

4.2 温水制冷存在的挑战

4.2.1 温水制冷

当前大部分数据中心处于温带或热带地区^②, 制冷能耗开销大, 因此温水制冷是一种有效的节能手段^[96,97]。图4-3展示了通过实验测量得到的 CPU 在不同利用率和

① Amazon EC2 Instance Types, <https://aws.amazon.com/ec2/instance-types/>

② Data Center Map, <https://www.datacentermap.com/>

不同温度制冷液下的温度变化 (CPU 型号: Intel Xeon E5-2650 V3, 能耗频率管理策略: “powersave”, 最大温度阈值: 78.9 摄氏度^①)。可以看到在温水制冷中, 当 CPU 处于负载较高时, 其温度接近设计的最大温度阈值。而使用 50 摄氏度的水制冷时, CPU 温度会超过最大温度阈值。值得注意的是, 当 CPU 利用率超过 50% 时, 其温度增长速度变缓慢。这是由于 CPU 默认使用了 Intel CPU 能耗频率管理驱动 *p_state* 提供的“powersave”策略, 当 CPU 利用率小于 50% 时, 其运行频率会随着利用率的增加, 从 1.2GHz 逐渐增加到 2.3GHz, 当 CPU 利用率超过 50% 时, CPU 的频率被锁定在 2.3GHz。因此在 CPU 利用率超过 50% 之后, 其温度的增长趋势变缓。CPU 长时间运

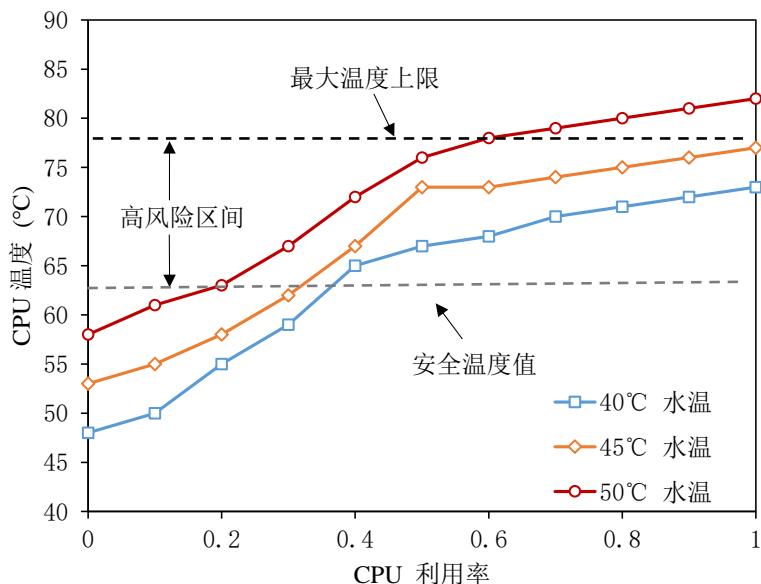


图 4-3 CPU 温度随利用率和水温的变化趋势

行在接近或超过最大温度阈值的范围, 会引起系统不稳定甚至缩短 CPU 使用寿命^②。通常会设定一个低于最大温度阈值的安全运行温度 (如最大温度阈值的 80%) 作为 CPU 温度上限。显然采用了温水制冷后, CPU 在负载较高时, 存在一定的风险。目前数据中心的平均利用率普遍较低 (如阿里云数据中心平均利用率仅为 26.32%), 因此采用 40 至 45 摄氏度, 甚至 50 摄氏度的温水进行制冷存在很大的可行空间。但是

① Intel Xeon E5-2650 v3 specifications, <http://www.cpu-world.com/CPUs/Xeon/Intel-Xeon%20E5-2650%20v3.html>

② Min./Max. Temperature, http://www.cpu-world.com/Glossary/M/Minimum_Maximum_operating_temperatures.html

如图4-4和4-5所示，当把谷歌^①和阿里巴巴^[98]公布的数据中心的利用率在时间和空间上进行分析后发现，数据中心负载不但存在4.1中描述的时间维度上的不均衡，还存在空间维度上的不均衡，存在局部的服务器利用率很高的现象。因此，温水制冷仍旧存在一定的风险。

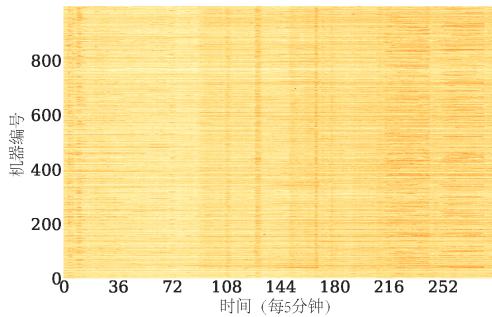


图 4-4 谷歌数据中心利用率

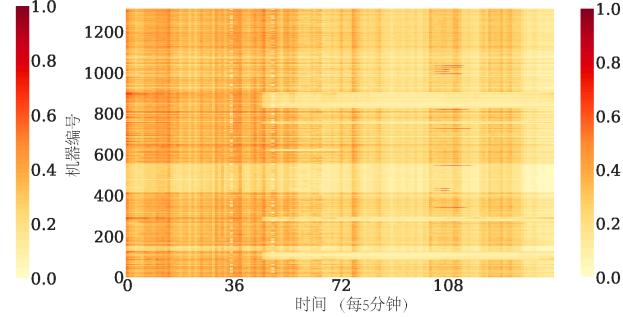


图 4-5 阿里巴巴数据中心利用率

CPU 利用率	温度 (°C)																			
	节点1		节点2		节点3		节点4		节点5		节点6		节点7		节点8		节点9		节点10	
CPU0	CPU1	CPU0	CPU1	CPU0	CPU1	CPU0	CPU1	CPU0	CPU1	CPU0	CPU1	CPU0	CPU1	CPU0	CPU1	CPU0	CPU1	CPU0	CPU1	
0%	31	31	32	35	33	32	32	31	32	33	32	32	31	33	32	31	30	29		
20%	35	35	35	38	34	33	37	36	33	34	36	35	34	34	35	35	32	31	30	
40%	38	38	36	41	38	36	39	39	35	36	38	37	36	36	37	37	35	35	33	
60%	39	39	39	44	40	37	42	41	37	38	40	39	38	38	39	38	37	37	35	
80%	39	39	40	46	40	38	43	42	38	38	40	40	38	39	40	38	39	38	38	
100%	40	40	41	48	42	40	44	44	38	39	42	42	40	40	41	40	41	39	40	

图 4-6 同一个水冷系统中不同节点 CPU 温度值（供水温度 20 摄氏度）

4.2.2 细粒度制冷控制

目前数据中心温度不均衡主要由两个因素造成。第一是前文提到的数据中心服务器之间的利用率不均衡。第二是由于水冷系统本身的制冷效率是不均衡的。图4-6展示了在同一个机柜中 10 个硬件配置相同的曙光 TC4600E-LP 服务器节点的温度测量结果。曙光 TC4600E-LP 服务器采用的是水冷系统，所有服务器 CPU 均由一个中心化的泵进行制冷液循环散热。可以看到，即使是每个 CPU 的利用率相同，它们的温度也各有差异。例如，在满载时，节点 2 的 CPU1 温度比节点 5 的 CPU0 高出

^① Google cluster workload traces, <https://github.com/google/cluster-data>

10 度。这主要由于不同服务器和泵的空间距离不同，造成制冷液水压和流速在不同，最终导致制冷效率不一致。

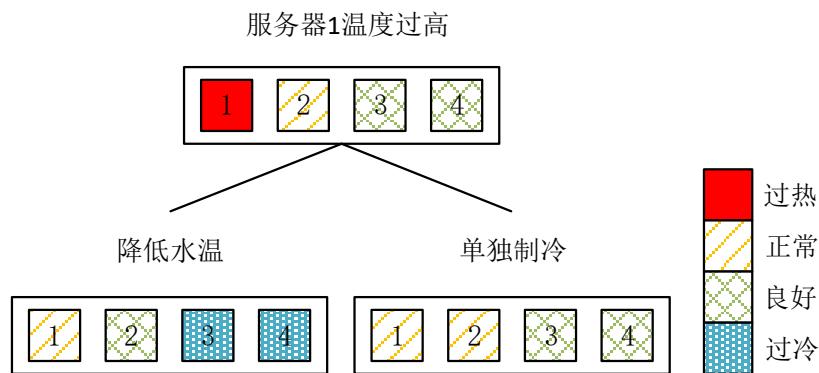


图 4-7 热点解决方法对比

在传统水冷系统中，负载不均衡并不会对水冷系统的效率造成任何问题。这是由于传统水冷系统采用冷水制冷的策略，存在大量冗余的制冷能力。即使是 CPU 利用率很高的情况下，其温度依旧很低。如图4-6测量结果所示，采用 20 摄氏度的制冷水的情况下，即使所有 CPU 的负载都达到了 100%，最高的 CPU 温度也仅为 48 摄氏度。但是在 40~45 甚至更高温度的温水制冷中，负载不均衡会导致局部热点，大大制约温水制冷的效率。如图4-7所示，服务器 1 过热需要进行降温。在当前水冷系统中，唯一的办法就是通过制冷机降低全局水温，将服务器 1, 2, 3, 4 同时降低温度。但是服务器 2, 3, 4 原本并没有过热，**对这些服务器制冷就会造成能耗浪费。显然更加有效率的办法是仅对服务器 1 提供单独额外的制冷，这就需要一种细粒度的制冷控制方法。**

在当前水冷系统中，部署分布式的制冷液循环系统（即为每一个服务器安装单独的循环泵）来控制制冷液的流速，可以达到细粒度制冷控制的目的。但是在水冷系统中，CPU 热量仍旧是靠热传导传递到水中，而在热量传递中，温差是影响导热效率的最主要因素。流速控制并不能使 CPU 和制冷水之间温差迅速增大，而且在工程上难以维护，会带来额外的成本开销和一系列液压问题。因此这种方案并不被水冷方案提供商所推荐采用^[88]。

4.3 混合水冷系统的设计

本节将介绍混合水冷系统的构造与设计，来解决温水制冷中存在的挑战。

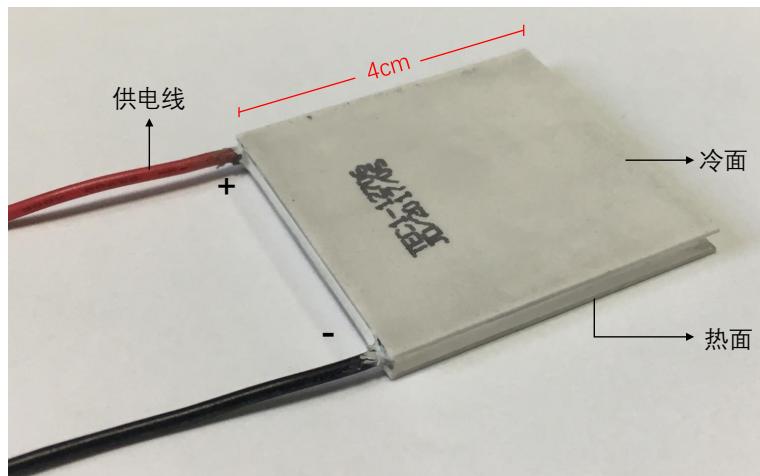


图 4-8 半导体制冷片

4.3.1 半导体制冷片的特点

图4-8展示了半导体制冷片（Thermoelectric Cooler, TEC）。TEC 的原理是热电效应，即在直流电的作用下，热能被从一面带到另外一面，从而产生冷面和热面。冷面进而可以用来进行制冷。

物理特性：TEC 最高可输入 12V 的直流电，而且可以通过调节直流电的电压来改变制冷能力。其电压需求和传统 CPU 风扇完全相同，因此可以直接通过服务器主板供电。TEC 的尺寸为 $4 \times 4\text{cm}$ ，和 CPU 的尺寸相同，因此冷面可以直接贴合 CPU 进行制冷。

成本分析：TEC 的价格非常便宜。以图4-8中型号为 TEC1-12706 为例，其价格仅为 10-15 元^①。其平均故障间隔时间（Mean Time Before Failure, MTBF）为 25 至 30 万小时^②，远远大于服务器的使用寿命，因此 TEC 是一种便宜、稳定的制冷部件，即使大规模应用到数据中心的成本也不高。

① 半导体制冷片价格，<https://m.tb.cn/h.e2XE2r1?sm=8db1fc>

② Thermoelectric Cooler Reliability Testing and Reports, <https://www.tec-microsystems.com/faq/reliability-testing-and-reports.html>

局限性：TEC 的制冷需要消耗电力，其峰值功耗约为 60 瓦。因此，使用 TEC 对 CPU 制冷时，需要根据 CPU 的温度来调节 TEC 的输入功率，来达到最佳的能效。此外，TEC 的工作原理是通过电场将热能从一面带到另外一面，热面的热量和温度会不断增加，如果热面的热量没有及时散出去，会导致 TEC 损坏。因此，在使用 TEC 为 CPU 进行制冷时，还需要考虑 TEC 本身的散热需求。TEC 本身在不工作时的导热性较差，无法作为导热介质，因此在整合 TEC 的同时，水冷系统的构造需要重新设计，从而实现 TEC 制冷和水冷之间的无缝切换。

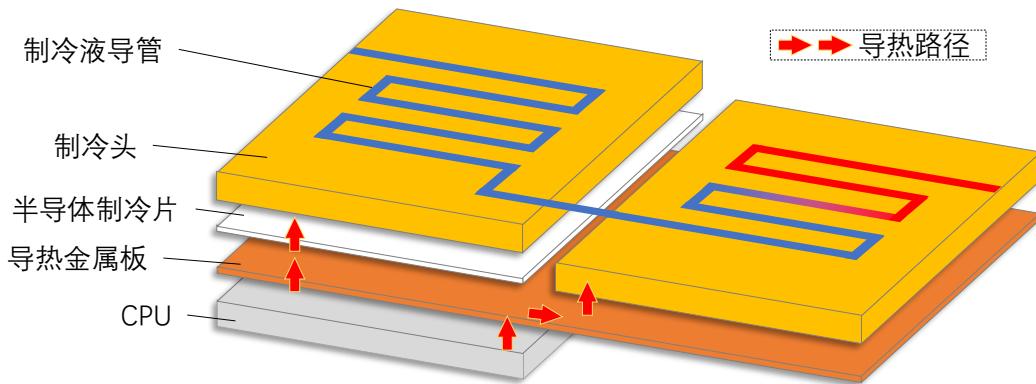


图 4-9 混合水冷结构设计

4.3.2 混合水冷结构设计

为了将 TEC 整合进现有的水冷系统并实现水冷和 TEC 制冷之间的无缝切换，本章提出了一种新的混合水冷结构，如图4-9所示。在传统的水冷系统中，每个 CPU 上贴合一个水冷头。CPU 的热量传导至水冷头，制冷液流过水冷头将热量带出服务器。本章提出的新的水冷结构增加了新的三个部件：TEC，导热金属板和额外的一个水冷头。当系统监测到 CPU 温度小于预先设定的安全温度阈值的时候，TEC 不工作。由于 TEC 在不通电工作时的导热性较差，CPU 热量通过铜片，传导到水冷头中，如图4-9中右侧的热量传导路径所示。当 CPU 负载升高或者水温过高无法有效制冷时，TEC 开始工作，并将热量传递另外一个水冷头中，如图4-9中左侧热量传导路径所示。通过上述新的硬件重构，可以实现传统水冷和 TEC 制冷之间的无缝切换。

将 TEC 和水冷头通过导热铜片相连来实现制冷手段的切换存在一个缺点。当

TEC 工作时，铜片的温度会比水温低。这会使得右侧水冷头中的热量传到回铜片上，影响 TEC 的制冷效果。尽管如此，根据下一小节的真实平台测量，TEC 仍旧能达到很高的制冷效率。

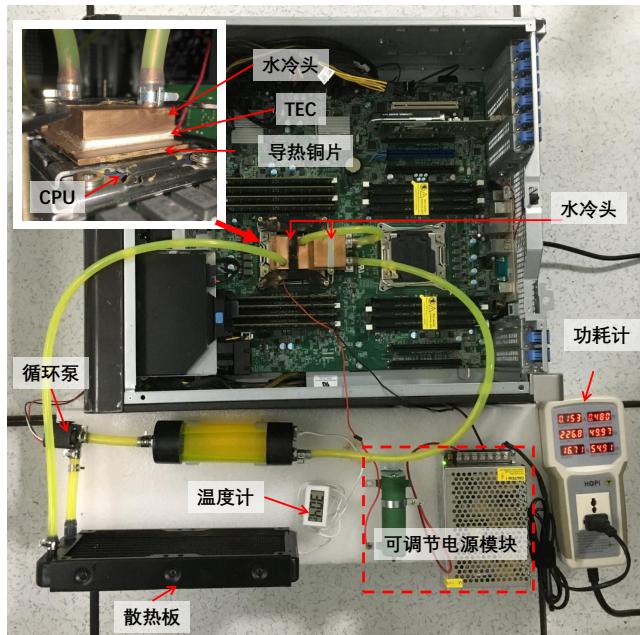


图 4-10 混合水冷系统原型

4.3.3 混合水冷系统实现与效率测评

在上述的混合水冷结构中，需要评估两个重要参数：TEC 的功耗和 TEC 的制冷能力。为了达到这一目的，本小节在服务器上搭建了一个真实的混合水冷系统平台，如图4-10所示。该服务器使用的是 Intel Xeon E5 2650 V3 CPU，默认的频率控制策略为“powersave”。图4-10中左上角的为 CPU 侧面放大图，其和上一小节的构造一致，从下到上依次为 CPU、导热铜片、TEC 和制冷头。左下角是制冷液的循环系统，包含循环泵、散热板和一个监控水温的温度计。为了测量 TEC 的功耗，TEC 由一个外部的可调节电源模块单独供电，并用功耗计进行能耗监控，如图4-10右下角所示。值得注意的是，除了上一小节提到的 TEC、导热铜片、TEC 和制冷头，其他的部件（如外部电源模块、温度计等）并不是混合水冷系统的一部分。这些额外部件只用于系统原型的评测。

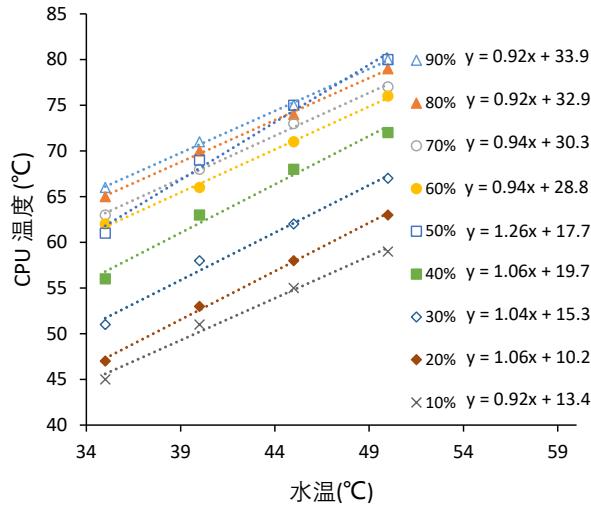


图 4-11 CPU 温度随着水温线性增长

实验首先评测了 TEC 不工作时，水冷的工作效率。值得注意的是，传统的水冷中 CPU 热量是直接传导到水冷头中，而在混合水冷系统中，CPU 的热量是经过导热铜片传递到水冷头中。但是通过实验的测量对比发现，这两种结构的水冷系统对 CPU 的降温效果没有区别。图4-11展示了在不同 CPU 利用率下，CPU 温度随水温的变化趋势。可以看到，当 CPU 利用率一定时，CPU 的温度随着水温的升高而升高，且呈现线性增长的趋势，且斜率约等于 1。这意味着在 TEC 不工作时，要将 CPU 温度降低 ΔT ，就需要将水温降低 ΔT 。

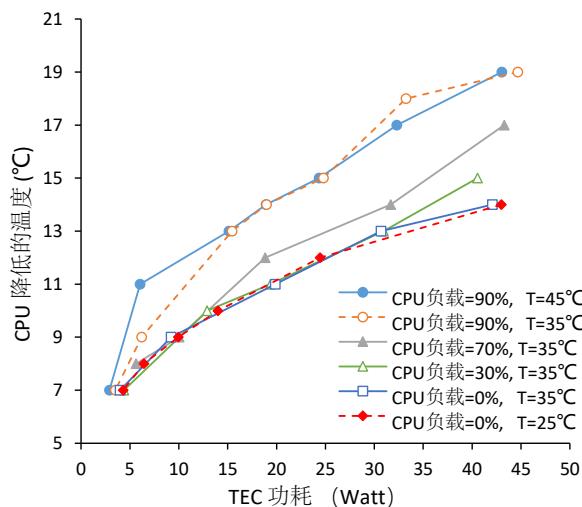


图 4-12 TEC 的制冷效率

实验然后测量了 TEC 的制冷能力和功耗。显然 TEC 的功耗越高，其制冷能力越强。为了进一步探究其他因素的 TEC 制冷能耗的影响，实验还测试了在不同 CPU 负载和不同水温下的 TEC 制冷能力，如图4-12所示（T 代表供水温度）。通过比较 CPU 负载 =90% 时，T=35 和 45 摄氏度的情况，以及 CPU 负载 =0% 时，T=25 和 35 摄氏度的情况可以看出，当 CPU 负载固定时，水温的变化对 TEC 制冷能力的影响很小。而 CPU 的负载对 TEC 的制冷能力有着明显的影响，而且当 CPU 负载越高，相同功率下 TEC 能够降低更多的温度。通过上述真实的测量可以验证，混合水冷系统中，TEC 能够有效降低 CPU 温度，而且可以通过功率输入进行降温能力的调节。

4.4 混合水冷系统的部署框架

本节将介绍如何在真实的数据中心部署混合水冷系统和制冷控制。

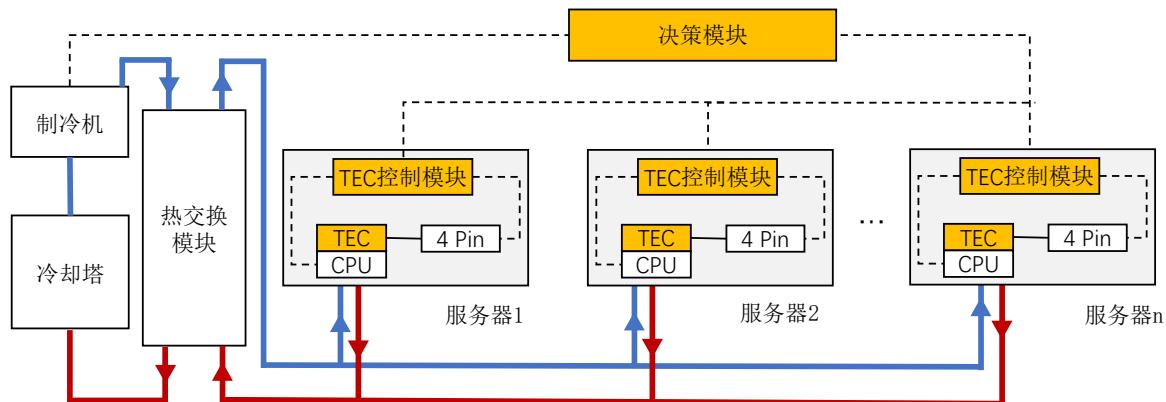


图 4-13 混合水冷系统原型

总览：图4-13展示了混合水冷系统在数据中心部署的框架图。和传统水冷系统一样，该框架包含了制冷机、冷却塔等模块。此外还包含了 TEC 的控制模块，根据不同服务器的功耗以及散热需求，动态调节 TEC 的工作状态。

TEC 控制模块：该模块通过 *lm_sensors* 软件实现，主要由两个功能：收集 CPU 温度信息和控制 TEC 的输入功率。由于 TEC 的输入电压和传统的 CPU 风扇相同，因此可以直接连接到服务器主板上的 4-pin 电源连接器。4-pin 电源连接器通过脉冲宽度调制（pulse-width modulation, PWM）来控制其输出功率，来改变风扇的转速。同理，

该功能也可以控制 TEC 的工作功率。具体而言，*lm_sensors* 软件提供了 *pwmconfig* 和 *fancontrol* 两个工具来设置功率输出策略，实现自动调整 4-pin 电源连接器输出功率。通过 *lm_sensors* 软件，可以实现对 TEC 制冷能力的控制。

决策模块：决策模块收集来自 TEC 控制模块的 CPU 温度信息，从而得到整个数据中心的温度图。然后根据温度图来决策控制制冷机和 TEC 的工作频率，实现混合水冷方法。显然决策模块的机制决定了混合水冷方案的整体能效。下面将介绍如何实现能效最优的混合水冷控制方法。

4.5 IT 负载感知的自适应混合水冷控制方法

本节将介绍混合水冷的控制方法。假设冷却塔至多能将水温降低至 $T_{WarmWater}$ ，用该水温的水对服务器进行制冷后，任意一个服务器 i 在 j 时刻的 CPU 温度为 T_{ij} 。如果该 CPU 在 j 时刻的温度超过预先设定的安全温度阈值 T_{safe} 时，则需要降低 $\Delta T = T_{ij} - T_{safe}$ 摄氏度。

显然数据中心的服务器负载是不断变化的，且在空间上呈现不均衡的状态。从图4-12可以看到，当 TEC 在需要降低更多的 CPU 温度时，其能效比（降低的温度/功耗）是呈现下降的趋势的。纯 TEC 的制冷方案并不能带来最佳的能效。当热点服务器的数量比例较大时，使用制冷机进行全局降温能效更高。因此采用制冷机和 TEC 混合的制冷方案可以获取最佳能效。这就需要解决如下两个问题：（1）何时启用制冷机制；（2）制冷机应该提供多少制冷量。这两者和数据中心的负载变化息息相关。因此，如何根据数据中的负载变化提供合理的混合制冷方案是一大挑战。

如果可以准确预测下一个时段的不同服务器的负载、功耗或者温度等信息，则可以通过计算并比较制冷机在提供不同制冷量时的整体能耗，来预先得到最佳设定。然而，负载预测是困难的，而且存在误差。尤其是在用户众多的公有云数据中心内，用户的任务提交具有很大的随机性和不可控性。

为了解决以上问题，本节提出一种自适应的混合水冷控制方法。对于每个时刻 j ，可以监测到每个 CPU 的温度^①，从而可以得到一张服务器的温度分布图，如图4-14所

^① CPU 的温度可以通过软件，如 *lm_sensors* 直接读取。

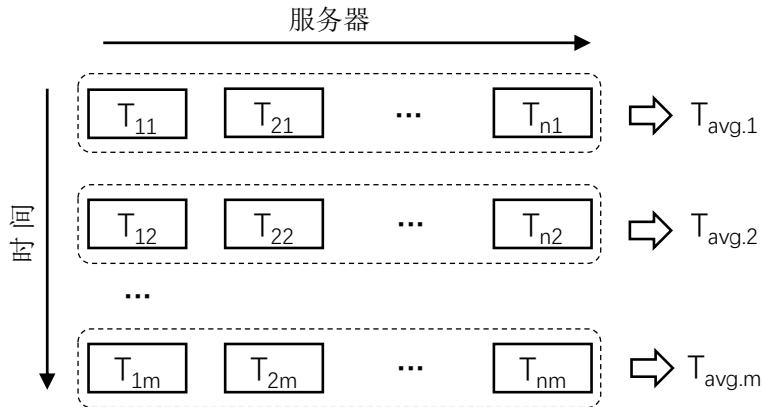


图 4-14 数据中心的温度图

示。其中 CPU 温度 T_{ij} 超过安全温度阈值 T_{safe} 的称为热点。

为了说明方便, 图4-15采用了 CPU 的平均温度 T_{avg} 来表示每个时刻的总体制冷需求。制冷机是通过降低水温来进行制冷, 其制冷是全局制冷, 因此只有当热点数量超过一个预先设定的热点比例阈值 P_{ct} (如 80%) 时, 才使用制冷机。具体而言, 如图4-15所示, 自适应的混合水冷控制方法按如下步骤处理温水制冷中的热点问题:

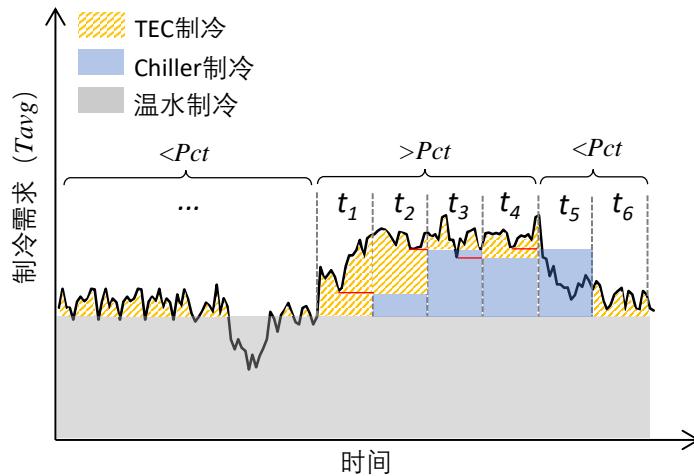


图 4-15 自适应的混合水冷控制

- **步骤一:** 在第一个出现热点比例高于 P_{ct} 的制冷控制周期时 (如 t_1), 只采用 TEC 作为制冷手段。TEC 可以提供实时的制冷需求响应, 避免延迟产生的制冷需求不匹配。

- **步骤二**: 在下一个制冷控制周期开始时(如 t_2), 如果热点的比例在前一个时间控制周期内(即 t_1) 大于 Pct , 那么制冷机将根据前一个时间控制周期内的最低的制冷需求提供制冷量, 即设定制冷机降低水温 $\Delta T = \min\{T_{avg,1}, T_{avg,2}, \dots, T_{avg,m}\} - T_{safe}$ 。否则, 转**步骤三**。
- **步骤三**: 如果热点的比例在前一个时间控制周期内(如 t_5) 小于 Pct , 那么在该时间控制周期内只用 TEC 进行制冷(如 t_6)。

本节提出的自适应混合水冷控制方法主要通过前一个制冷控制周期的制冷需求, 来制定下一个周期内的制冷策略。通过不断重复上述的步骤, 就能根据服务器的负载变化, 实现动态的制冷控制方法, 同时无需负载预测等复杂手段, 也更加具有通用性。

4.6 性能测评

本节将具体介绍实验设置和评估混合水冷系统及其控制方法的能效。

4.6.1 实验设置

由于缺乏真实的混合水冷的数据中心, 本章将通过真实的谷歌和阿里巴巴数据中心的服务器负载数据, 来进行仿真实验。

CPU 温度模型: 为了实验方便, 本章假设数据中心的服务器是同构的, 即服务器都采用了4.3节中的 Intel Xeon E5-2650 V3 型号的处理器, 且 CPU 能耗频率管理策略都采用了 Intel 提供的“powersave”。根据图4-3的测量结果, CPU 的温度 T_{CPU} 和其利用率 u 之间的关系可以用线性分段函数进行表示为:

$$T_{CPU}(u) = \begin{cases} 36.29u + 57.09, & \text{if } u \in [0, 0.5], T_{water} = 50^{\circ}\text{C} \\ 10u + 72, & \text{if } u \in [0.5, 1], T_{water} = 50^{\circ}\text{C} \end{cases} \quad (4.1)$$

异构数据中心由于采用的 CPU 不同, 因此功耗以及温度也会不同。但是本章提出的制冷控制方法是基于 CPU 的温度的, 而 CPU 温度可以通过系统工具(如 *lm_sensors* 软件) 获取, 所以本章提出的制冷控制方法同样适用于异构数据中心。

区别在于数据中心温度图的数值不同。实验后续将采用不同的数据中心服务器利用率数据集，从而产生不同的温度图来评测本章提出的解决方案。

在另一方面，如图4-11所示，CPU的温度随着水温呈线性增长的关系，可以近似表示为：

$$T_{CPU} = T_{water} + D(u) \quad (4.2)$$

其中 $D(u)$ 是 CPU 利用率为 u 时，线性函数在图4-11中的截距。

通过上述两个温度模型，可以根据数据中心服务器的使用率数据，来得出数据中心不同时刻各个服务器的 CPU 温度。

制冷机能耗模型：假设在时间 t 内，一部分的服务器 CPU 温度值超过了安全温度阈值，其中温度最高的服务器超过安全温度阈值 ΔT 。为了使所有服务器温度回到安全值以下，制冷机需要在时间 t 内将流入的水温至少降低 ΔT 。假设流速为 F ，单位为 m^3/h ，那么在时间 t 内，制冷机消耗的能耗为：

$$E_{ConventionalWaterCooling} = C_{water} * \Delta T * F * t * \rho / COP_{chiller}. \quad (4.3)$$

其中 $C_{water} = 4.2 \times 10^3 \text{ J/(kg}\cdot\text{C)}$ 代表了水的比热容，即 1 千克的水降低 1 摄氏度所需要移除的热量为 4.2×10^3 。 $F * t$ 代表了在时间 t 内制冷机需要降低温度的水的体积， ρ 代表了水的密度。 $COP_{chiller}$ (Coefficient of Performance) 是热力学中描述制冷机能效比的参数，即单位功率下的制冷量 ($COP = \text{制冷机移除的热量}/\text{制冷机消耗的能量}$)。

TEC 能耗模型：当使用 TEC 来应对上述相同场景下的服务器制冷需求时，可以对每个服务器单独提供制冷。从图4-12的测量结果可知，TEC 的效率受到 CPU 负载的影响。为了不失一般性，本章实验采用性能最差的曲线来描述 TEC 的功耗，即图4-12中 CPU 负载 =0%，T=25°C 的曲线。为了更好地拟合该曲线，实验采用了分段的二次函数来进行表达：

$$P_{TEC}(\Delta T) = \begin{cases} 0.092\Delta T^2 + 0.034\Delta T, & \text{if } \Delta T \in [0, 8] \\ 0.89\Delta T^2 - 14.16\Delta T + 65.43, & \text{if } \Delta T \in (8, 14] \end{cases} \quad (4.4)$$

假设在于任意一个时刻, N 台服务器超过了安全温度阈值, 且分别超出 $\Delta T_1, \Delta T_2, \dots, \Delta T_N$ 。那么 TEC 所消耗的制冷功耗可以表示为:

$$E_{TEC} = \sum_{i=1,\dots,N} P_{TEC}(\Delta T_i) \quad (4.5)$$

值得说明的是, 以上的模型在真实的数据中心部署时, 并不需要。例如 CPU 的温度可以直接通过系统工具监测, 制冷机的能耗可以直接用功率计测得。对于 TEC, 本章采用了其性能最差的功耗模型, 来保证能耗评测的有效性和可靠性。

对比设置: 为了更好地评测混合水冷系统设计, 实验采用了三种不同类型的数据中心服务器使用率数据集进行评估, 并设置了三个对比组进行能效对比。

波动型: 该数据由阿里巴巴数据中心^①的 1313 台服务器在 12 小时内的服务器使用率构成。其总体的服务器使用率呈现较为剧烈和频繁的波动变化。

不规则型: 谷歌数据中心^②提供了超过 12 万台服务器在一个月内的使用率数据。为了简化, 实验从中挑选了 1000 台服务器在 24 小时内的使用率组成实验数据。该实验数据总体 CPU 利用率除了存在较小的利用率高峰, 还伴随突发的较大使用率高峰。

常规型: 类似地, 实验从谷歌提供的数据中另外挑选了 1000 台服务器在 24 小时内的使用率组成新的实验数据。该实验数据的总体使用率较为平缓, 不存在突发的大的使用率高峰。

最优对比组 (hybridOpt): 如果可以准确预测服务器的负载、功耗、温度等信息, 就可以提前计算并制定最优的混合制冷策略。由于是采用已知的数据中心数据集进行实验, 所以可以根据各个时刻的服务器温度, 直接结算出最优方案和能耗作为对照组。

制冷机对比组 (chiller): 对于传统的水冷系统, 如果实施了温水制冷且可以容忍制冷延迟, 则可以使用制冷机来为超温的服务器进行制冷。其特点是需要根据温度最高的服务器来设定制冷机的制冷量, 其能耗可以根据式 (4.3) 计算得出。

^① Alibaba cluster workload traces, <https://github.com/alibaba/clusterdata>

^② Google cluster workload traces, <https://github.com/google/cluster-data>

TEC 对比组 (*TEC*)：另外可以不使用制冷机，而直接使用 *TEC* 来为超温的服务进行降温，特点是可以为每个服务器进行单独制冷，其能耗可以根据式 (4.5) 计算得出。

表 4.1 混合水冷实验中其他参数的设置

参数	设置
制冷机能效比 ($COP_{chiller}$)	3.6
水流速 (F)	$0.5m^3/h$
安全温度阈值 (T_{safe})	63°C

其他设置：表4.1总结了实验中的其他参数设置，其依据如下：

- 制冷机 COP ($COP_{chiller}$)：根据制冷机制造标准，制冷的能效比最低标准为 $2.7 \sim 3.6$ ^[99]。实验中将 COP 设置为 3.6。
- 水流速 (F)：在4.2节中测量的 Sugon TC4600ELP 服务器，其水流速度设置 $0.5m^3/h$ 。因此实验参考真实系统设置，将流速设置为相同值。
- 安全温度阈值 (T_{safe})：CPU 长时间处于接近或者超过最大温度阈值的范围时，会造成系统的不稳定甚至缩短 CPU 寿命。为了保证温水制冷中系统的稳定性，实验将安全温度阈值设置为最大温度阈值的 80%。对应地，实验中采用的 Intel Xeon-E5-2650 V3 CPU 的安全温度阈值为 63°C。

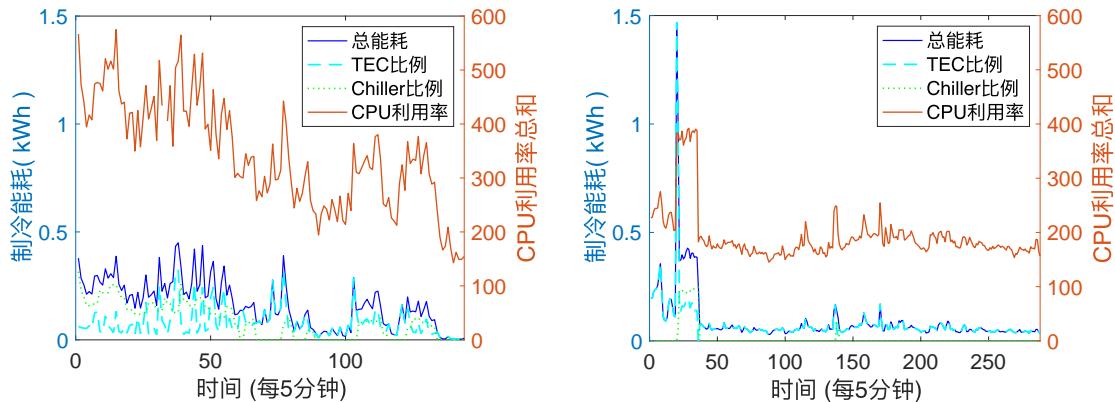


图 4-16 波动型负载中的混合制冷能耗特征

图 4-17 不规则型负载中的混合制冷能耗特征

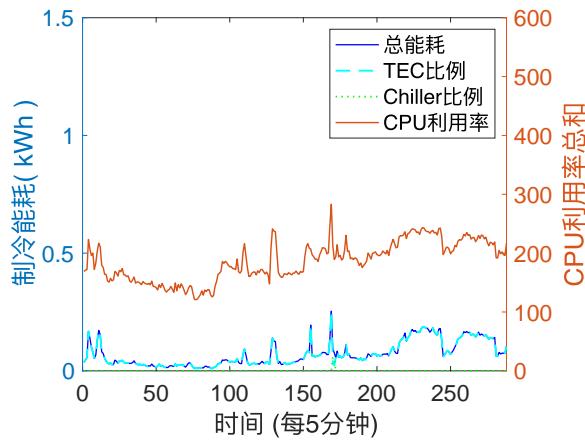


图 4-18 常规型负载中的混合制冷能耗特征

4.6.2 混合水冷系统的能耗模式分析

图4-16至图4-18展示了实验中采用的三种不同类型（波动型，不规则型，常规型）的数据中心服务器负载，以及本文提出的混合制冷控制方法中制冷机和TEC的能耗情况（实验设置： $T_{water} = 50^{\circ}\text{C}$, $P_{ct} = 80\%$, 制冷控制周期为15分钟）。可以看到，总制冷能耗是随服务器的总负载的变化趋势而变化的，这避免了不必要的制冷和能耗浪费。其中在波动型的制冷能耗中，制冷机的工作比例比不规则型和常规型要高。这是由于在波动型的数据中总体的CPU利用率更高，总体的CPU温度也更高，更容易触发制冷机工作。在不规则型和常规型负载中，制冷机除了在CPU利用率急剧升高时，其能耗大部分时候为0。这是由于这两种数据类型中，CPU的总体利用率较低，热点的比例在大部分时候比实验设定的触发比例80%要低，因此制冷机没有工作。

4.6.3 不同制冷策略的能效对比

图4-19展示了不同制冷策略的能耗对比结果（ $T_{water} = 50^{\circ}\text{C}$, $P_{ct} = 80\%$, 制冷控制周期为15分钟）。和制冷机对比组的能耗（ $E_{chiller}$ ）相比，本章提出的混合控制方法的能耗（ E_{hybrid} ）在波动型、不规则型和常规型数据中，分别降低了58.72%，74.48%，78.43%的能耗。

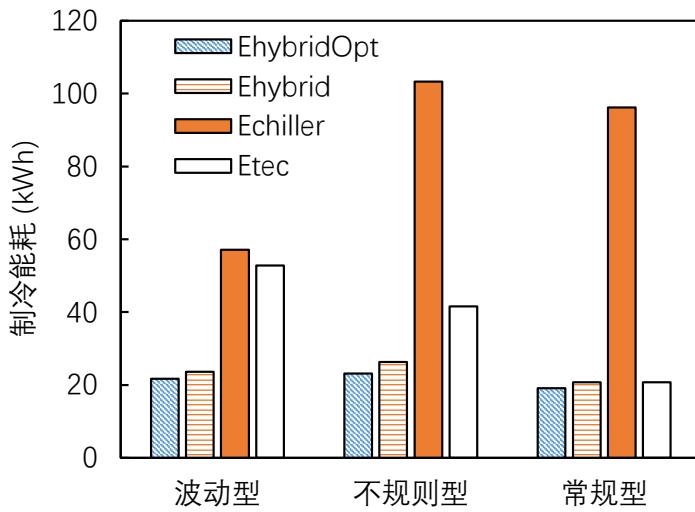


图 4-19 不同策略的制冷能耗对比

通常 CPU 的能耗和其利用率之间存在线性关系，即两者的关系可以用线性函数进行表达^[22]。实验中采用的 Intel Xeon-E5-2650 V3 CPU 的最高设计功耗为 105 W。三种类型的数据中平均 CPU 利用率分别 26.32%，19.62%，18.40%，服务器数量分别为 1313, 1000, 1000，运行时间为 12 小时，24 小时，24 小时。因此可以粗略计算出三种类型数据中 CPU 消耗的总电力分别为 435.43 kWh, 494.42 kWh, 436.68 kWh。本章根据 CPU 的 IT 能耗和制冷能耗，定义 CPU 局部能效比（partial Power Usage Effectiveness, pPUE）：

$$pPUE = \frac{CPU\ Energy + CPU\ Cooling\ Energy}{CPU\ Energy} \quad (4.6)$$

那么根据图4-19的制冷能耗结果可知，当采用 50 摄氏度的温水制冷时，制冷机对比组在三种负载中的 pPUE 分别能达到 1.12, 1.21 和 1.21，而本文提出的混合制冷系统可以达到 1.05, 1.05 和 1.04。

4.6.4 不同参数设置对混合制冷控制方法能效的影响

下面将讨论不同设置（水温 T_{water} 、制冷控制周期和热点比例阈值 P_{ct} ）对本文提出的混合控制方法能耗的影响。

冷却塔水温：冷却塔的水温取决于外部空气温度，炎热的天气意味着更高的水

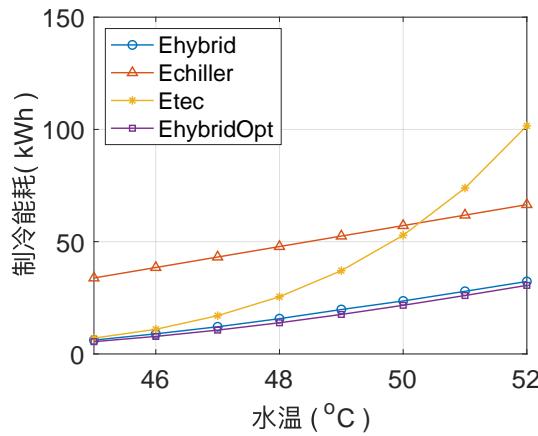


图 4-20 水温对波动型负载制冷能耗的影响

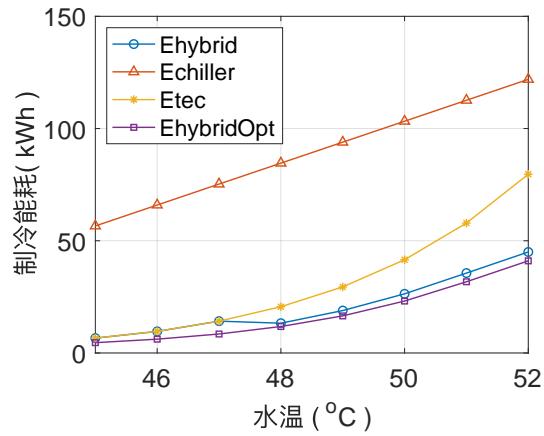


图 4-21 水温对不规则型负载制冷能耗的影响

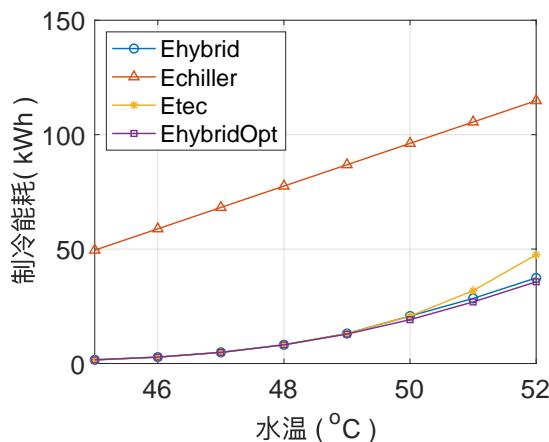


图 4-22 水温对常规型负载制冷能耗的影响

温。图4-20至图4-22显示了来自冷却塔的水温增长时的制冷能耗比较 ($P_{ct} = 80\%$, 制冷控制周期为 15 分钟)。由于冷却塔的降温能力降低, 水的温度增加, 因此每个 CPU 的温度也会升高, 增加了制冷机/TEC 的使用频率。在图4-20中, E_{TEC} 变得比 $E_{chiller}$ 大, 这是因为实验采用了 TEC 的最差制冷能力的函数进行仿真, 即当降低更多的温度时, TEC 的效率 (降低的温度/消耗的功率) 变得更差。从图4-21中可以看出当水温低于 48 摄氏度时, E_{hybrid} 和 E_{TEC} 是相同的。这是因为热点比例低于实验设置的阈值 0.8, 制冷机不工作, 而当水温高于 48 摄氏度时, 热点比例增加, 并触发混合冷却, 从而实现更好的冷却效率。

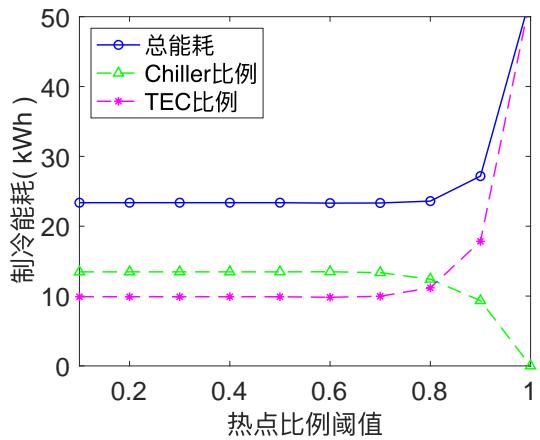


图 4-23 热点比例阈值对波动型负载制冷能耗的影响

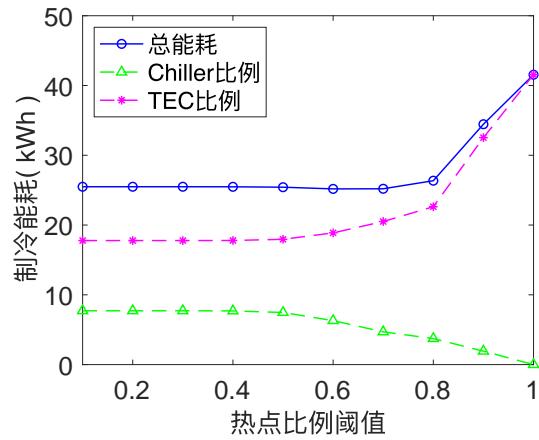


图 4-24 热点比例阈值对不规则型负载制冷能耗的影响

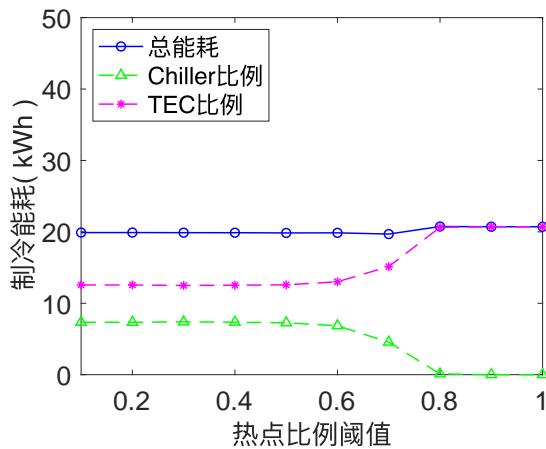


图 4-25 热点比例阈值对常规型负载制冷能耗的影响

热点比例阈值：图4-23至图4-25显示了不同热点比例阈值设置对应的能量消耗 ($T_{water} = 50^{\circ}\text{C}$, 制冷控制周期为 15 分钟)。阈值设置大于 0.8 时将导致更多的制冷能耗。这是因为过高的热点比例阈值设置很难激活混合制冷。可以看到，0.8 是不同工作负载的合适上限阈值，可以实现较好的制冷效率。阈值变小时，能量消耗仅略微增加。这是因为当热点的百分比较低时，其平均 CPU 温度较低。本章提出的混合制冷控制方法中，制冷机提供的冷却能力是基于最低的平均 CPU 温度，因此即使制冷被激活，制冷机提供的冷却能力也很小。通过这种方式，可以避免过度冷却一些服务

器和能源浪费。这使本文提出的混合制冷方法对于小的热点比例阈值设置更加具有鲁棒性。

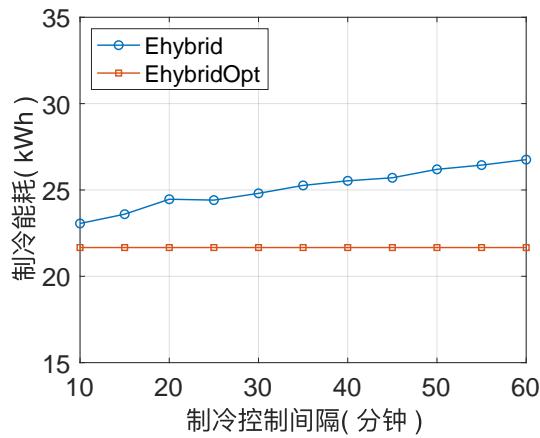


图 4-26 制冷控制周期对波动型负载制冷能耗的影响

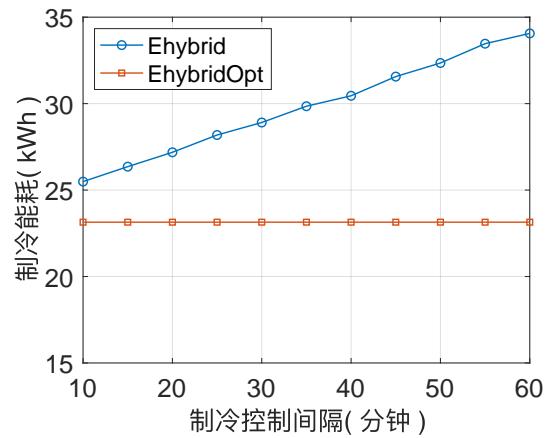


图 4-27 制冷控制周期对不规则型负载制冷能耗的影响

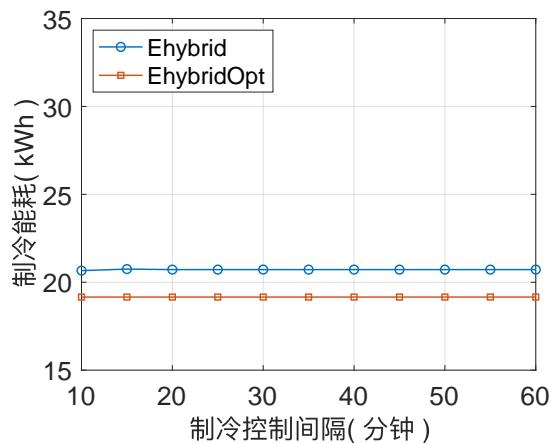


图 4-28 制冷控制周期对常规型负载制冷能耗的影响

制冷控制周期：这是另一个可能影响混合制冷能效的参数。图4-26至图4-28显示了制冷控制间隔设置对应的制冷能耗 ($T_{water} = 50^{\circ}\text{C}$, $P_{ct} = 80\%$)。可以发现，较小的制冷控制间隔可以达到更好的制冷能效。这是因为本章提出的制冷控制方法是基于前一个制冷控制间隔的冷却需求，设置较小的间隔可以更频繁地更新制冷策略，从而更快地适应制冷需求的变化。但是，如果可以预测制冷需求，那么制冷控制间

隔对制冷能效没有影响（如图4-26至图4-28中 $E_{hybridOpt}$ 所示）。特别地，可以看到时间控制间隔对波动型和不规则型的数据中心负载有更大的影响。这是因为这两个数据中心负载都具有显着的利用率变化。从图4-18可以看出，在常规型数据中心负载中，只有一小段时间会触发混合冷却。因此冷却控制间隔对它几乎没有影响。总的来说，较小的冷却控制间隔可以在本章提出的混合制冷方法中达到更好的制冷能效。

4.7 本章小结

通过对数据中心 IT 负载的研究发现，数据中心平均 IT 负载（即服务器的利用率）普遍较低，造成了制冷能力的冗余和能耗浪费。此外在时间和空间上存在不均衡的现象，导致数据中心在时空上存在散热需求的差异化。本章基于数据中心负载较低的特点并利用冗余的制冷能力所提供的调节空间，提出使用温水制冷来减少制冷能耗。为了应对温水制冷带来制冷失败的风险，设计了一种基于半导体制冷片的混合水冷系统，该系统可以为原本的水冷系统提供细粒度的制冷能力。为了进一步优化制冷能效。另外系统还将 IT 负载变化和制冷控制相结合，设计了一种自适应的混合制冷控制方法。通过真实的数据中心负载数据的仿真测试发现，本章提出的混合水冷系统可以达到 1.04~1.05 的局部能效比，大大降低了数据中心的能耗成本。

5 总结与展望

随着云计算的迅速发展，数据中心作为云计算的底层支撑设施其规模也急剧扩张。数据中心的规模增长给数据中心运营和管理带来诸多挑战，其中能耗开销给数据中心的运行成本和云服务的定价造成了巨大压力。因此优化数据中心能效，降低能耗成本对数据中心发展尤为关键。能耗优化的第一步是通过能耗计量来理解数据中心的能耗行为。数据中心作为一个复杂的基础设施，不仅包含服务器等 IT 设施，还包含了电力系统、制冷系统等非 IT 设施。IT 设施和非 IT 设施是数据中心能耗最主要的两种来源。因此能耗计量和优化需要同时面向 IT 和非 IT 两个层面。

针对数据中心能耗计量和优化两个方面，本文基于经济学方法设计了面向 IT 虚拟层和非 IT 设施层的能耗计量方法，并针对数据中心 IT 负载不均衡的现象，设计了高能效的制冷系统，主要创新成果具体如下：

(1) 面向 IT 虚拟机层的能耗计量方法

针对解决虚拟机资源竞争导致的能耗难以计量的问题，本文通过将虚拟机的能耗计量和经济学利益分配方法中的夏普利值结合，提出了一种具有理论支撑的能耗计量方法来保证结果的公平性。针对虚拟机能耗在运行过程中的变化，本文通过引入虚拟机状态对传统的夏普利值进行了动态扩展，设计了动态夏普利值方法，其次针对动态夏普利值输入数量繁多而且虚拟机配置不同的问题，进一步设计了线性的虚拟机集合体能耗评估方法，实现了公平的虚拟机能耗计量方法。

(2) 面向非 IT 设施层的能耗计量方法

针对目前非 TI 能耗计量的空白，本文设计了一种基于夏普利值的轻量级计量方法。非 IT 设施能耗在物理层面难以进行直接划分，因此也可以利用夏普利值方法进行细粒度的能耗计量。然而夏普利值的计算复杂度高达 $O(2^N)$ ，而一个非 IT 设施通常服务成千上万的 IT 设备，因此夏普利值方法难以直接应用到非 IT 能耗的计量。本文测量并分析真实数据中心非 IT 设施的能耗，并基于非 IT 设施能耗特征进行推导证明，设计了一种和夏普利值方法等价但是复杂度仅为 $O(N)$ 的方法：非 IT 设施的动态能耗按虚拟机 IT 能耗比例进行分配，非 IT 设施的静态能耗在虚拟机之间均分。

该方法既保证了非 IT 能耗计量的公平性，又降低了计算复杂度。

(3) 基于 IT 负载感知的混合水冷系统

为了提高数据中心的能效比，本文设计了一种高能效的混合水冷系统。数据中心负载和功耗在时间和空间上存在不均衡的现象，导致制冷系统能效低下。目前水冷系统由于技术限制，采用的都是中心化的制冷控制方法，导致制冷效率低下。本文通过将和半导体制冷片和水冷系统相结合，设计了新型的水冷系统结构，实现了数据中心细粒度的制冷控制。为了进一步优化制冷能效，本文采用温水制冷，并根据数据中心负载变化设计了制冷机和半导体制冷片相结合方法：在服务器温度普遍较高时，启用制冷机的全局制冷和半导体制冷片的细粒度制冷，在服务器温度普遍较低时，只采用半导体制冷片制冷。该策略能够为服务器提供精确的制冷，避免制冷能耗的浪费。

最后，在数据中心能耗计量和优化方面，本文认为还存在如下一些研究问题值得进一步探究：

(1) 虚拟机能耗评价标准。虚拟机之间存在资源竞争的现象使得传统资源-能耗映射模型无法有效量化资源竞争对能耗的影响。虚拟机本质是一种软件，没有物理实体，其能耗无法通过硬件直接测量。这导致了软件方法获取的虚拟机能耗结果无法得到直接验证。因此虚拟机的能耗评价标准无法以准确性作为评价指标。本文尝试从经济学方法出发，提出了基于公平性的能耗计量方法，然而公平性的能耗计量结果并不等于准确的虚拟机能耗。因此，如何获取每个虚拟机的准确能耗仍然是数据中心能耗计量的一个挑战。

(2) 基于机器学习的数据中心节能。鉴于数据中心结构的复杂性和监测数据的丰富性，机器学习非常适合数据中心环境。现代大型数据中心具有各种机械和电气设备，以及对应的设置和控制方案。这些系统与各种反馈回路之间的相互作用使得使用传统工程公式难以准确地预测数据中心效率。例如，冷通道温度设定的简单改变将在冷却设施（如冷却器，冷却塔，热交换器，泵等）中产生负载变化，这反过来导致设备效率的非线性变化，此外环境天气条件也会影响最终的制冷效率。使用标准公式进行预测建模通常会产生较大误差，因为它们无法捕获这种复杂的相互依赖性。如何应用机器学习来对数据中心能耗行为建模，并给出合理的能效优化指导建议也

是一个值得研究的问题。

(3) 数据中心热量回收。数据中心在消耗大量电力的同时，也产生大量的热量。随着智慧城市概念的发展，数据中心已成为处理智慧城市产生的大量数据的技术基础。因此，许多数据中心位于城市环境中，为实施各种智慧城市服务提供支持。降低数据中心能耗解决方案的一种方法是将数据中心与智能城市的能源基础设施和公用事业集成，使数据中心连接到供暖网络中。IT 部件产生的废热可以有效地重复使用，用于供暖等。废热再利用将成为数据中心的一个相当大的财务收入流，有效降低能耗成本的压力。因此，如何将数据中心废热再利用也是一个值得研究的问题。

致 谢

从 2014 年进入实验室到现在，转眼已经过去了 5 年。在这里，我不仅学到了关于数据中心的前沿知识，还掌握了做研究的方法，使自己的专业水平有了很大提升，人生也有了新的方向。

首先要感谢的是我的导师金海教授。金老师领导的 CGCL 实验室给我们创建了资源丰富的科研平台和资助环境。他虽然工作繁忙，但还是坚持按时和我们开博士沙龙，以渊博的专业知识和广阔的视野为我们的科研进展指导方向、及时解决科研上的困难。在本文的写作上，金老师也给予了细心的指导。在平时交流中也能感受到金老师的平易近人和对学生的关系，而且金老师总是把学生的利益放在第一位，学生的任何问题在金老师这里都是大事情，总是能第一时间帮助解决。非常谢谢您一直以来的关怀和帮助！

同时还要特别感谢我的指导老师刘方明教授。刘老师在我的博士生涯中给了无数的帮助。犹记得我第一次接触科研时非常迷茫，是刘老师细心的指导让我在科研上找到了合适的方向。也记得第一次参与论文写作时刘老师细心指导写作的正确方法。正是在刘老师小组中一次又一次磨砺当中，使我掌握了大量的研究方法和经验。

另外还要感谢来自加拿大维多利亚大学的吴奎教授、加州大学石溪分校的任少磊教授、目前在国防科技大学任职的唐国明师兄以及帮助一起做实验、修改论文的冯思锐学弟、贾子阳学弟、朱心慧学妹，感谢他们百忙之中抽空在我论文上给予巨大帮助。同时要感谢实验室的各位老师，包括王多强老师、刘英书老师、耿聪老师等其他所有老师，谢谢你们为实验室创造了良好的工作环境，使我们能够在一个轻松、自在的环境中进行科研工作。

五年的读博生活，我还认识了一群共同奋斗的同窗好友。郭鉴学长、易小萌学长、周知学长、唐皓文学长、许志峰学长、吴广原学长，感谢你们在我科研工作上的指点和帮助。还有和我一起入学汪涛、费新财、牛轶佩、陈姝彤等同学，大家一起共同奋斗，见证了互相的成长，感谢你们一路的陪伴和帮助，还有小组的学弟学妹们，谢谢你们的共同陪伴，让我的研究生生涯更加丰富多彩。

华 中 科 技 大 学 博 士 学 位 论 文

感谢我的爸爸、妈妈，感谢你们对我一直以来的支持和鼓励，深深地感谢我的姐姐，在我远离家乡的日子里照顾父母以及在我学业和生活上的关心与支持。你们是我在人生道路上一直奋斗的最大动力。

最后，对所有关心和帮助过我的人表示最衷心的感谢！

参考文献

- [1] Kavis M J. Architecting the cloud: design decisions for cloud computing service models (SaaS, PaaS, and IaaS). John Wiley & Sons, 2014.
- [2] 开放数据中心委员会. 数据中心白皮书 (2018) . 2018.
- [3] Idex G. Cisco global cloud index: Forecast and methodology, 2016–2021,. Cisco, San Jose, CA, USA, White Paper C11-738085-02, Şubat, 2018.
- [4] Koomey J. Growth in data center electricity use 2005 to 2010. A report by Analytical Press, completed at the request of The New York Times, 2011, 9.
- [5] Greenberg A, Hamilton J, Maltz D A, et al. The cost of a cloud: research problems in data center networks. ACM SIGCOMM computer communication review, 2008, 39(1):68–73.
- [6] Avgerinou M, Bertoldi P, Castellazzi L. Trends in data centre energy consumption under the European code of conduct for data centre energy efficiency. Energies, 2017, 10(10):1470.
- [7] Whitehead B, Andrews D, Shah A, et al. Assessing the environmental impact of data centres part 1: Background, energy use and metrics. Building and Environment, 2014, 82:151–159.
- [8] 邓维, 刘方明, 金海, 等. 云计算数据中心的新能源应用: 研究现状与趋势. 计算机学报, 2013.
- [9] Cosmano J. Choosing a data center. Disaster Recovery Journal, 2012.
- [10] Ashrae T. 9.9 (2011). Whitepaper prepared by ASHRAE technical committee (TC), 2011, 9.
- [11] Evans T. Fundamental principles of air conditioners for information technology. White Paper, 2004, 57:2004–1.

- [12] Popa P. Managing server energy consumption using IBM PowerExecutive. IBM Systems and Technology Group, Tech. Rep, 2006.
- [13] 罗亮, 吴文峻, 张飞. 面向云计算数据中心的能耗建模方法. 软件学报, 2014.
- [14] 林伟伟, 吴文泰. 面向云计算环境的能耗测量和管理方法. 计算机学报, 2016.
- [15] Isci C, Martonosi M. Runtime power monitoring in high-end processors: methodology and empirical data. in: Proceedings of the International Symposium on Microarchitecture (MICRO). IEEE, 2003, 93-104.
- [16] Rotem E, Naveh A, Ananthakrishnan A, et al. Power-management architecture of the intel microarchitecture code-named sandy bridge. in: Proceedings of the International Symposium on Microarchitecture (MICRO). IEEE, 2012, 20–27.
- [17] Huang W, Lefurgy C, Kuk W, et al. Accurate Fine-Grained Processor Power Proxies. in: Proceedings of the International Symposium on Microarchitecture (MICRO). IEEE, 2012, 224–234.
- [18] Fan X, Weber W D, Barroso L A. Power Provisioning for a Warehouse-sized Computer. in: Proceedings of the International Symposium on Computer Architecture (ISCA). ACM, 2007, 13–23.
- [19] Lim M Y, Porterfield A, Fowler R. SoftPower: fine-grain power estimations using performance counters. in: Proceedings of the International Symposium on High Performance Distributed Computing (HPDC). ACM, 2010, 308–311.
- [20] Ranganathan P, Leech P, Irwin D, et al. Ensemble-level Power Management for Dense Blade Servers. in: Proceedings of the International Symposium on Computer Architecture (ISCA). IEEE, 2006, 66–77.
- [21] Rivoire S, Ranganathan P, Kozyrakis C. A Comparison of High-Level Full-System Power Models. in: Proceedings of the Workshop on Power-Aware Computing and Systems (HotPower). USENIX, 2008, 32–39.

- [22] Tang G, Jiang W, Xu Z, et al. Zero-Cost, Fine-Grained Power Monitoring of Datacenters Using Non-Intrusive Power Disaggregation. in: Proceedings of the Annual Middleware Conference (MIDDLEWARE). ACM, 2015, 271–282.
- [23] Tang G, Jiang W, Xu Z, et al. NIPD: Non-intrusive power disaggregation in legacy datacenters. IEEE Transactions on Computers (ToC), 2017, 66(2):312–325.
- [24] Tsirogiannis D, Harizopoulos S, Shah M A. Analyzing the Energy Efficiency of a Database Server. in: Proceedings of the International Conference on Management of Data (SIGMOD). ACM, 2010, 231–242.
- [25] Kansal A, Zhao F. Fine-grained Energy Profiling for Power-aware Application Design. in: Proceedings of the International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2008.
- [26] Koller R, Verma A, Neogi A. WattApp: An Application Aware Power Meter for Shared Data Centers. in: Proceedings of the International Conference on Autonomic Computing (ICAC). ACM, 2010, 31–40.
- [27] Zhai Y, Zhang X, Eranian S, et al. HaPPy: Hyperthread-aware Power Profiling Dynamically. in: Proceedings of the Annual Technical Conference (ATC). USENIX, 2014, 211–217.
- [28] Kansal A, Zhao F, Liu J, et al. Virtual Machine Power Metering and Provisioning. in: Proceedings of the International Symposium on Cloud Computing (SoCC). ACM, 2010, 39–50.
- [29] Krishnan B, Amur H, Gavrilovska A, et al. VM Power Metering: Feasibility and Challenges. in: Proceedings of the International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2011, 56–60.
- [30] Bohra A E H, Chaudhary V. VMeter: Power modelling for virtualized clouds. in: Proceedings of the International Symposium on Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010, 1-8.

- [31] Ben-Yehuda M, Day M D, Dubitzky Z, et al. The Turtles Project: Design and Implementation of Nested Virtualization. in: Proceedings of the International Symposium on Operating Systems Design and Implementation (OSDI). USENIX, 2010, 423–436.
- [32] Colmant M, Kurpicz M, Felber P, et al. Process-level Power Estimation in VM-based Systems. in: Proceedings of the European Conference on Computer Systems (EuroSys). ACM, 2015, 14:1–14:14.
- [33] 叶可江, 吴朝晖, 姜晓红, 等. 虚拟化云计算平台的能耗管理. 计算机学报, 2012.
- [34] 武晋, 何利力. 云计算数据中心能耗优化研究综述. 软件导刊, 2019.
- [35] Andrew L L, Lin M, Wierman A. Optimality, Fairness, and Robustness in Speed Scaling Designs. in: Proceedings of the International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2010, 37–48.
- [36] Huang M, Renau J, Yoo S M, et al. A framework for dynamic energy efficiency and temperature management. in: Proceedings of the International Symposium on Microarchitecture (MICRO). ACM, 2000, 202–213.
- [37] David H, Fallin C, Gorbatov E, et al. Memory power management via dynamic voltage/frequency scaling. in: Proceedings of the International Conference on Autonomic Computing (ICAC). ACM, 2011, 31–40.
- [38] Gurumurthi S, Sivasubramaniam A, Kandemir M, et al. DRPM: dynamic speed control for power management in server class disks. in: Proceedings of the International Symposium on Computer Architecture (ISCA). IEEE, 2003, 169–179.
- [39] Liu Y, Draper S C, Kim N S. SleepScale: runtime joint speed scaling and sleep states management for power efficient data centers. in: Proceedings of the International Symposium on Computer Architecture (ISCA). IEEE, 2014, 313–324.
- [40] Somu Muthukaruppan T, Pathania A, Mitra T. Price theory based power management for heterogeneous multi-cores. in: Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). ACM, 2014, 161–176.

- [41] Meisner D, Gold B T, Wenisch T F. PowerNap: eliminating server idle power. in: Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). ACM, 2009, 205–216.
- [42] Sudan K, Srinivasan S, Balasubramonian R, et al. Optimizing Datacenter Power with Memory System Levers for Guaranteed Quality-of-service. in: Proceedings of the International Conference on Parallel Architectures and Compilation Techniques (PACT). ACM, 2012, 117–126.
- [43] Pinheiro E, Bianchini R. Energy Conservation Techniques for Disk Array-based Servers. in: Proceedings of the International Conference on Supercomputing (ICS). ACM, 2004, 68–78.
- [44] Liu Z, Lin M, Wierman A, et al. Greening geographical load balancing. in: Proceedings of the International Conference on Measurement and modeling of computer systems (SIGMETRICS). ACM, 2011, 233–244.
- [45] Zhang Y, Wang Y, Wang X. Greenware: Greening cloud-scale data centers to maximize the use of renewable energy. in: Proceedings of ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing. Springer, 2011, 143–164.
- [46] Zhou Z, Liu F, Zou R, et al. Carbon-aware online control of geo-distributed cloud services. IEEE Transactions on Parallel and Distributed Systems (TPDS), 2016, 27(9):2506–2519.
- [47] Kontorinis V, Zhang L E, Aksanli B, et al. Managing Distributed Ups Energy for Effective Power Capping in Data Centers. in: Proceedings of the International Symposium on Computer Architecture (ISCA). IEEE, 2012, 488–499.
- [48] Liu L, Li C, Sun H, et al. HEB: Deploying and Managing Hybrid Energy Buffers for Improving Datacenter Efficiency and Economy. in: Proceedings of the International Symposium on Computer Architecture (ISCA). ACM, 2015, 463–475.

- [49] Wang D, Ren C, Sivasubramaniam A. Virtualizing Power Distribution in Datacenters. in: Proceedings of the International Symposium on Computer Architecture (ISCA). ACM, 2013, 595–606.
- [50] Hsu C H, Deng Q, Mars J, et al. SmoothOperator: Reducing Power Fragmentation and Improving Power Utilization in Large-scale Datacenters. in: Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). ACM, 2018, 535–548.
- [51] Khatamifard S K, Wang L, Yu W, et al. ThermoGater: Thermally-aware on-chip voltage regulation. in: Proceedings of the International Symposium on Computer Architecture (ISCA). IEEE, 2017, 120–132.
- [52] Agrawal A, Torrellas J, Idgunji S. Xylem: enhancing vertical thermal conduction in 3D processor-memory stacks. in: Proceedings of the International Symposium on Microarchitecture (MICRO). ACM, 2017, 546–559.
- [53] Beigi M V, Memik G. Thermal-aware optimizations of reRAM-based neuromorphic computing systems. in: Proceedings of the Annual Design Automation Conference (DAC). ACM, 2018, 39.
- [54] 李翔, 姜晓红, 吴朝晖, 等. 绿色数据中心的热量管理方法研究. 计算机学报, 2015.
- [55] Moore J D, Chase J S, Ranganathan P, et al. Making Scheduling “Cool”: Temperature-Aware Workload Placement in Data Centers. in: Proceedings of the USENIX Annual Technical Conference (ATC), 2005, 61–75.
- [56] Bash C, Forman G. Cool Job Allocation: Measuring the Power Savings of Placing Jobs at Cooling-Efficient Locations in the Data Center. in: Proceedings of the USENIX Annual Technical Conference (ATC), 2007, 140.

- [57] Le K, Bianchini R, Zhang J, et al. Reducing electricity cost through virtual machine placement in high performance computing clouds. in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC). ACM, 2011, 22:1–22:12.
- [58] Skach M, Arora M, Hsu C H, et al. Thermal Time Shifting: Leveraging Phase Change Materials to Reduce Cooling Costs in Warehouse-scale Computers. in: Proceedings of the International Symposium on Computer Architecture (ISCA). ACM, 2015, 439–449.
- [59] Skach M, Arora M, Tullsen D, et al. Virtual Melting Temperature: Managing Server Load to Minimize Cooling Overhead with Phase Change Materials. in: Proceedings of the International Symposium on Computer Architecture (ISCA). IEEE, 2018, 15–28.
- [60] Manousakis I, Goiri I n, Sankar S, et al. CoolProvision: Underprovisioning Datacenter Cooling. in: Proceedings of the International Symposium on Cloud Computing (SoCC). ACM, 2015, 356–367.
- [61] Goiri I n, Nguyen T D, Bianchini R. CoolAir: Temperature- and Variation-Aware Management for Free-Cooled Datacenters. in: Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). ACM, 2015, 253–265.
- [62] Delforge P. America’s data centers are wasting huge amounts of energy. Natural Resources Defense Council (NRDC), 2014, pages 1–5.
- [63] Islam M A, Mahmud H, Ren S, et al. Paying to save: Reducing cost of colocation data center via rewards. in: Proceedings of the International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2015, 235–245.
- [64] Dines R A. Build or buy? the economics of data center facilities. Forrester Research, 2011.
- [65] Bircher W L, John L K. Complete System Power Estimation Using Processor Performance Events. IEEE Transactions on Computers (ToC), 2012, 61(4):563–577.

- [66] 薛利敏. 夏普利值在利益分配中的应用. 商场现代化, 2006.
- [67] 易欣, 张飞涟, 邱慧. 不确定 AHP 和 Shapley 值应用于投标联合体利益分配. 计算机工程与应用, 2012.
- [68] Shapley L S. A value for n-person games. Contributions to the Theory of Games, 1953, 2(28):307–317.
- [69] Misra V, Ioannidis S, Chaintreau A, et al. Incentivizing Peer-assisted Services: A Fluid Shapley Value Approach. in: Proceedings of the International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2010, 215–226.
- [70] Stanojevic R, Laoutaris N, Rodriguez P. On Economic Heavy Hitters: Shapley Value Analysis of 95Th-percentile Pricing. in: Proceedings of the International Conference on Internet Measurement (IMC). ACM, 2010, 75–80.
- [71] Dong M, Lan T, Zhong L. Rethink Energy Accounting with Cooperative Game Theory. in: Proceedings of the International Conference on Mobile Computing and Networking (MobiCom). ACM, 2014, 531–542.
- [72] Amazon still lags behind Apple, Google in Greenpeace renewable energy report. <https://www.greenpeace.org/archive-international/en/press/releases/2017/Amazon-still-lags-behind-Apple-Google-in-Greenpeace-renewable-energy-report/>.
- [73] Jiang W, Liu F, Tang G, et al. Virtual machine power accounting with shapley value. in: Proceedings of the International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017, 1683–1693.
- [74] Feng X, Ge R, Cameron K W. Power and energy profiling of scientific applications on distributed systems. in: Proceedings of the International Symposium on Parallel and Distributed Processing Symposium (IPDPS). IEEE, 2005, 10–pp.
- [75] Avelar V, Azevedo D, French A, et al. PUE: a comprehensive examination of the metric. White paper, 2012, 49.

- [76] Rack DCLC Product Guide. <http://coolit2017.qt4egaquh7.maxcdn-edge.com/wp-content/uploads/2017/09/rack-dclc-product-guide.pdf>.
- [77] Du Su Y L. GreenMap: mapreduce with ultra high efficiency power delivery. in: Proceedings of the Workshop on Hot Topics in Cloud Computing (HotCloud). USENIX, 2015, 6–6.
- [78] Sawyer R L. Making large UPS systems more efficient. Electron Journal-South African Institute of Electrical Engineers, 2006, 23(6):65.
- [79] Liu Z, Chen Y, Bash C, et al. Renewable and Cooling Aware Workload Management for Sustainable Data Centers. in: Proceedings of the International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2012, 175–186.
- [80] Pelley S, Meisner D, Wenisch T F, et al. Understanding and abstracting total data center power. in: Proceedings of the Workshop on Energy-Efficient Design (WEED), 2009.
- [81] Zhang Q, Shi W. UPS-aware workload placement in enterprise data centers. IEEE Computer Magazine, 2015.
- [82] Gopalakrishnan R, Marden J R, Wierman A. Potential games are necessary to ensure pure nash equilibria in cost sharing games. Mathematics of Operations Research, 2014, 39(4):1252–1296.
- [83] Islam M A, Ren S, Wierman A. Exploiting a Thermal Side Channel for Power Attacks in Multi-Tenant Data Centers. in: Proceedings of the International Conference on Computer and Communications Security (CCS). ACM, 2017, 1079–1094.
- [84] Wang C, Urgaonkar B, Gupta A, et al. Effective Capacity Modulation as an Explicit Control Knob for Public Cloud Profitability. in: Proceedings of the International Conference on Autonomic Computing (ICAC). IEEE, 2016, 95-104.
- [85] 诸凯, 刘泽宽, 何为, 等. 数据中心服务器 CPU 水冷散热器的优化设计. 制冷学报, 2019.

- [86] Frizziero M. Rethinking Chilled Water Temperatures Can Bring Big Savings in Data Center Cooling. <https://blog.schneider-electric.com/datacenter/2016/08/17/water-temperatures-data-center-cooling/>, 2016.
- [87] Shan Y, Huang Y, Chen Y, et al. LegoOS: A Disseminated, Distributed OS for Hardware Resource Disaggregation. in: Proceedings of the International Symposium on Operating Systems Design and Implementation (OSDI). USENIX, 2018, 69–87.
- [88] CoolIT. Centralized vs. Distributed Pumping for Rack-based Direct Liquid Cooling. https://www.coolitsystems.com/wp-content/uploads/2017/07/coolit_centralized_vs_distributed_pumping_rev08.pdf, 2017.
- [89] Kang K D, Alian M, Kim D, et al. VIP: Virtual Performance-State for Efficient Power Management of Virtual Machines. in: Proceedings of the International Symposium on Cloud Computing (SoCC). ACM, 2018, 237–248.
- [90] Chaudhry M T, Ling T C, Manzoor A, et al. Thermal-aware scheduling in green data centers. ACM Computing Surveys (CSUR), 2015, 47(3):39.
- [91] Lo D, Cheng L, Govindaraju R, et al. Towards Energy Proportionality for Large-Scale Latency-Critical Workloads. in: Proceedings of the International Symposium on Computer Architecture (ISCA). IEEE, 2014, 301–312.
- [92] Zhang K, Guliani A, Ogreni-Memik S, et al. Machine Learning-Based Temperature Prediction for Runtime Thermal Management Across System Components. IEEE Transactions on Parallel and Distributed Systems (TPDS), 2018, 29(2):405–419.
- [93] Mars J, Tang L, Hundt R, et al. Bubble-up: Increasing utilization in modern warehouse scale computers via sensible co-locations. in: Proceedings of the International Symposium on Microarchitecture (MICRO). ACM, 2011, 248–259.
- [94] Steiner M, Gaglanello B G, Gurbani V, et al. Network-aware Service Placement in a Distributed Cloud Environment. in: Proceedings of the International Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). ACM, 2012, 73–74.

华 中 科 技 大 学 博 士 学 位 论 文

- [95] Alizadeh M, Edsall T, Dharmapurikar S, et al. CONGA: Distributed Congestion-aware Load Balancing for Datacenters. in: Proceedings of the International Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). ACM, 2014, 503–514.
- [96] Coles H, Ellsworth M, Martinez D J. "Hot" for Warm Water Cooling. in: Proceedings of State of the Practice Reports. ACM, 2011.
- [97] Conficoni C, Bartolini A, Tilli A, et al. Energy-aware Cooling for Hot-water Cooled Supercomputers. in: Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE). EDA Consortium, 2015, 1353–1358.
- [98] Lu C, Ye K, Xu G, et al. Imbalance in the cloud: an analysis on Alibaba cluster trace. in: Proceedings of the International Conference on Big Data(Big Data). IEEE, 2017, 2884–2892.
- [99] LLC P K. COPs, EERs, and SEERs How Efficient is Your Air Conditioning System? https://www.powerknot.com/wp-content/uploads/sites/6/2011/03/Power_Knot_about_COP_EER_SEER.pdf.

附录 1 缩略词简表

缩写	全称	中文释义
IaaS	Infrastructure-as-a-Service	基础设施即服务
PaaS	Platform-as-a-Service	平台即服务
SaaS	Software-as-a-Service	软件即服务
ICT	Information and Communication Technology	信息和通信技术
VM	Virtual Machine	虚拟机
CPU	Central Processing Unit	中央处理器
IT	Information Technology	信息技术
CDF	Cumulative Distribution Function	累积分布函数
CAGR	Compound Annual Growth Rate	复合年增长率
UPS	Uninterruptible Power Supply	不间断电源
PDU	Power Distribution Unit	电力分配单元
EC2	Amazon Elastic Computing Cloud	亚马逊弹性计算云
IDC	Internet Data Center	互联网数据中心
OAC	Outside Air Cooling	室外空气制冷
CRAC	Computer Room Air Conditioner	机房空调制冷机
PUE	Power Usage Effectiveness	电源使用效率
RAPL	Running Average Power Limit	运行功耗控制技术
HTT	Hyper Threading Technology	超线程技术
DVFS	Dynamic Voltage and Frequency Scaling	动态电压与频率控制技术
TEC	Thermoelectric Cooler	半导体制冷片
COP	Coefficient of Performance	能效比
PPUE	Partial Power Usage Effectiveness	局部能效比

附录 2 攻读博士学位期间发表论文

- [1] **Weixiang Jiang**, Fangming Liu, Guoming Tang, Kui Wu and Hai Jin, “Virtual Machine Power Accounting with Shapley Value”, in Proc. of IEEE International Conference on Distributed Computing System (ICDCS), Atlanta, Georgia, USA, June 5-8, 2017. (CCF B 类)
- [2] **Weixiang Jiang**, Shaolei Ren, Fangming Liu, Hai Jin, “Non-IT Energy Accounting in Virtualized Datacenter”, in Proc. of IEEE International Conference on Distributed Computing System (ICDCS), Vienna, Austria, July 2–5, 2018. (CCF B 类)
- [3] **Weixiang Jiang**, Ziyang Jia, Sirui Feng, Fangming Liu, Hai Jin, “Fine-grained Warm Water Cooling for Improving Datacenter Economy”, in Proc. of IEEE/ACM International Symposium on Computer Architecture (ISCA), Phoenix, Arizona, USA, June 22-26, 2019. (已接收录用, CCF A 类)
- [4] Guoming Tang, **Weixiang Jiang**, Zhifeng Xu, Fangming Liu, and Kui Wu, “Zero-Cost, Fine-Grained Power Monitoring of Datacenters Using Non-Intrusive Power Disaggregation”, in Proc. of ACM/IFIP/USENIX Middleare Conference (MIDDLEWARE), Vancouver, Canada, December 7-11, 2015. (CCF B 类)
- [5] Guoming Tang, **Weixiang Jiang**, Zhifeng Xu, Fangming Liu, Kui Wu, “NIPD: Non-Intrusive Power Disaggregation in Legacy Datacenters”, IEEE Transactions on Computers (ToC), 2016. (CCF A 类)
- [6] Chaobing Zeng, Fangming Liu, Shutong Chen, **Weixiang Jiang**, Miao Li, “Demystifying the Performance Interference of Co-located Virtual Network Functions”, in Proc. of IEEE International Conference on Computer Communications (INFOCOM), Honolulu, HI, USA, April 15-19, 2018. (CCF A 类)

附录 3 攻读博士学位期间参与的主要科研项目

- [1] 大型数据中心的低能耗可扩展理论与关键技术。国家自然科学重点项目。项目编号：No. 61133006。2012-2016。（已结题）
- [2] 软件定义的云数据中心网络基础理论与关键技术。国家重点基础研究发展计划 973 青年科学家项目。项目编号：No.2014CB347800。2014-2018。（已结题）
- [3] 高效能云计算数据中心资源调度关键技术研究。国家重点研发计划课题。项目编号：No.2017YFB1001703。2017-2021。（在研）

附录 4 个人简历

基本信息

姓 名：姜炜祥 性 别：男
院 校：华中科技大学计算机学院 专 业：计算机系统结构
电话号码：13517199760 政治面貌：群众
E-mail：wxjiang0905@gmail.com 导 师：金海 教授

研究兴趣

数据中心能耗

教育背景

2014.09-今	华中科技大学	计算机系统结构	博士
2010.09-2014.06	华中科技大学	信息安全	本科

研究经历

2017.12-2018.12	基于 IT 负载感知的混合水冷系统
2016.09-2017.12	面向非 IT 设施层的能耗计量方法
2015.08-2016.07	面向 IT 虚拟机层的能耗计量方法
2014.01-2015.12	非侵入式细粒度服务器能耗分解方法

所获奖励

2017.10	获华中科技大学计算机学院知行奖学金
2015.12	获华中科技大学计算机学院三好研究生称号
2015.09-2019.06	获华中科技大学研究生博士学业奖学金