数据中心跨层能效优化研究进展和发展趋势

王继业¹, 周碧玉^{2*}, 刘万涛², 虎嵩林²

1. 国家电网公司, 北京 100031

2. 中国科学院信息工程研究所, 北京 100089

* 通信作者. E-mail: zhoubiyu@iie.ac.cn

收稿日期: 2018-11-01; 接受日期: 2019-03-24; 网络出版日期: 2020-01-08

国家重点研发计划 (批准号: 2017YFB1010001) 和国家自然科学基金 (批准号: 61520206005, 61761136014) 资助项目

摘要 能耗问题已经成为阻碍数据中心进一步发展的主要障碍之一. 为了解决这个问题, 近年来, 学术界和工业界涌现了大量的数据中心能效优化研究. 本文针对数据中心节能问题, 首次从跨层能效优化的角度对国内外的研究现状进行分类和介绍, 包括基于 IT 负载调度的制冷系统能效优化、基于 IT 负载调度的供电系统能效优化, 以及 IT 系统与制冷系统跨层统一能效优化 3 个方面, 并展望了数据中心跨层能效优化研究的发展趋势.

关键词 能耗, 数据中心, 能效算法, 资源管理, 绿色计算

1 引言

云计算在全球范围内的蓬勃发展带动数据中心数目和规模的快速增长, 随之而来的是数据中心巨大的能耗. 美国资源保护协会预计, 2020 年全美数据中心的总功耗将高达 1400 亿千瓦时^[1]. 这意味着数据中心的电力支出将与购买新服务器的支出不相上下^[2]. 巨大的能耗开销已经成为阻碍数据中心进一步发展的绊脚石.

数据中心是一个层次化的系统, 主要由负责计算的 IT 系统层 (包括服务器、网络等)、负责为 IT 系统的安全高效运行提供保障的配套设施层 (包括制冷系统、供电系统、照明系统等), 以及基础设施层 (包括机房、地板等) 组成. 其中 IT 系统层和配套设施层是数据中心的两个能耗大户. 目前, 已有的数据中心节能研究绝大多数都是单独针对 IT 系统或者配套设施进行的. 例如, 文献 [3,4] 总结了数据中心的 IT 系统能效优化研究, 文献 [4~6] 总结了数据中心制冷系统和供电系统能效研究, 文献 [7] 总结了数据中心各个系统的能耗模型构建研究. 然而, 单一层面的能效优化可能会导致数据中心整体能效低下, 原因有两点: (1) 单纯优化计算系统能效的研究通常在配套设备供应有足够冗余的假设下进行, 这将导致巨大的配套设施能耗浪费; (2) 配套设施需要满足 IT 系统的需求, 因此, IT 系统的负

引用格式: 王继业, 周碧玉, 刘万涛, 等. 数据中心跨层能效优化研究进展和发展趋势. 中国科学: 信息科学, 2020, 50: 1-24, doi: 10.1360/N112018-00293

Wang J Y, Zhou B Y, Liu W T, et al. Research progress and development trend of cross-layer energy efficiency optimization in data centers (in Chinese). Sci Sin Inform, 2020, 50: 1-24, doi: 10.1360/N112018-00293

载大小与分布会决定配套设施的最小资源供应量. 而单纯优化配套设施能效的研究由于未与 IT 设备相互配合, 其节能空间有限. 综上所述, 将 IT 系统层与配套设施层联合考虑才能最大限度挖掘数据中心节能潜力, 实现提升数据中心整体能效的目的. 遗憾的是, 目前尚未有对数据中心跨层能效优化研究进行综述的文献.

本文首次从跨层能效优化的角度研究数据中心节能问题, 后续章节的组织如下: 第 2 节对数据中心跨层能效优化进行定义, 随后根据定义对国内外已有的相关研究进行分类; 第 3 节分别从基于 IT 负载调度的制冷系统能效优化、基于 IT 负载调度的供电系统能效优化和 IT 设备与制冷系统的跨层统一能效优化方案 3 个方面对数据中心跨层能效优化的最新研究进展进行分类总结和比较; 最后对全文进行总结, 并提出该领域未来的研究方向.

2 数据中心跨层能效优化

2.1 跨层能效优化的定义与分类

通常能效 (energy efficiency) 指的是电效率 (electric efficiency), 即一个电子电气工程中实体 (设备、组件或系统) 的有用的功率输出与所消耗的总功率的比值^[8]. 对于数据中心而言, 其有用功率输出主要表现在给用户提供的各种计算服务上面. 然而, 在实际数据中心内量化输出的计算服务量是非常困难的, 而且不利于在不同数据中心间进行能效比较. 目前, 人们普遍采用输入到 IT 设备的总功率与数据中心的总输入功率的比值来近似衡量数据中心的能效, 即 PUE (power usage effectiveness)^[9]. 从 PUE 的定义可以看出, PUE 的制定依据是数据中心的输入功率中只有一部分用于提供计算服务 (服务器、存储、网络等), 而剩下的功率则消耗在配套的支撑系统 (制冷、供电等) 中. 由于 IT 设备大部分情况下均未满载运行 (甚至负载普遍低下), 因此, PUE 在衡量数据中心的实际能效上仍然具有一定的局限性. 理想状态下, PUE 所能达到的最佳值是 1, 即输入数据中心的所有功率全部输入到内部的 IT 设备中. Google 宣称, 借助于自然冷源和可再生能源等措施, 他们的部分数据中心站点的 PUE 值能达到 1.06^[10]. 国内也普遍将 PUE 作为数据中心的能效评估标准, 例如 2018 年 9 月份出台的《北京市新增产业的禁止和限制目录 (2018 年版)》明确规定全市禁止新建和扩建 PUE 值高于 1.4 的数据中心¹⁾.

目前, 数据中心内耗能系统主要分布在 IT 系统层 (主要包括服务器和网络) 以及配套设施层 (主要包括制冷系统和供电系统). 虽然在不同的数据中心内, 上述系统的具体能耗占比值不同, 然而大体可以按照耗能占比值由高到低依次排序为服务器系统、制冷系统、供电系统、网络系统 (通常, 对于 PUE 较低的数据中心内, 服务器系统能耗占比最高, 而在 PUE 较高的数据中心内, 制冷系统能耗占比最高)^[4]. Pelley 等^[11]的一份研究报告表明, 在一个典型的数据中心内, 上述 4 个系统的能耗占比分别为 56%, 30%, 8% 和 5%. 结合 PUE 的定义不难看出, 提高数据中心能效的关键在于提高 IT 系统层的资源利用率, 同时降低配套设施层能耗.

跨层能效优化指的是结合上述两个层的耗能特性来进行综合能效优化, 通常是利用 IT 负载的调度实现, 即通过调整 IT 负载 (IT workload) 在 IT 设备上的分布、配合 IT 设备工作状态与供电制冷系统的参数配置, 达到在保障服务性能的前提下降低数据中心总能耗的目的. 依据 IT 系统是否与配套设施协同优化又可分为基于 IT 负载调度的 IT 系统层能效优化以及 IT 系统层与配套设施层的跨层统一能效优化两种 (如表 1 所示). 前者优化的目标仅仅是 IT 系统层的能效, 而 IT 负载调度仅仅是

1) <http://zhengce.beijing.gov.cn/library/192/33/50/438650/1566810/index.html>.

表 1 数据中心跨层能效优化研究分类
Table 1 Classification of cross-layer energy efficiency optimization in data center

Classification	System		
	IT	Cooling	Power distribution
Cooling optimization via IT scheduling	–	✓	–
Power distribution optimization via IT scheduling	–	–	✓
Joint optimization of IT and cooling	✓	✓	–
Joint optimization of IT and power distribution	✓	–	✓

手段; 后者 IT 系统与配套设施层两者的能效构建成统一的优化目标, 通过 IT 负载的调度、IT 设备工作状态的调整以及配套设施层的参数设置达到总能效最优化的目的. 目前已有的跨层研究主要为 IT 和制冷统一优化、IT 和供电统一优化以及 IT 和供电制冷系统统一优化 3 个方面. 本文也将从上述 3 个方面进行综述.

2.2 跨层统一能效优化的必要性

实施跨层统一的能效优化方案对于提高数据中心总体能效是非常必要的, 可以通过下面两个例子简单说明: (1) 仅仅针对 IT 系统层的能效优化方案通常将 IT 负载聚合到少数 IT 设备上, 通过尽可能多地关闭空闲 IT 设备以节省 IT 系统能耗. 然而, 该方案将不可避免在数据中心内产生若干热点. 为了保障 IT 设备安全运行, 配套设施层的制冷系统需要进一步调低冷气供应温度 (和/或提高风速) 以提供足够冷量来转移热点的热量, 进而产生更多的配套设施层能耗开销; 反过来, (2) 为了降低配套设施层的制冷系统能耗需要将 IT 负载尽可能均摊到所有 IT 设备上, 而此举又将提高 IT 设备的总能耗. 显然, 上述两种针对单层的节能措施并不能最终保证降低数据中心总体能耗. 文献 [12~16] 通过实验证明, 跨层统一的 IT 系统和制冷系统能效优化的研究方案比单一层面的能效优化方案可以节省 4%~30% 的能耗. 上述实验结果也印证了设计跨层统一的能效优化方案对提高数据中心总能效的必要性.

综上所述, 提出跨层统一的能效优化方案对于数据中心总能效的提升有着非常重要的意义. 一个可行的研究思路是综合配套设施层和 IT 系统层的设备特性对各层能耗进行定量刻画, 构建出跨层统一的能效优化模型, 并相应地提出跨层统一能效优化方案.

3 跨层能效优化研究进展

3.1 基于 IT 负载调度的制冷系统能效优化

数据中心内的 IT 设备在运行过程中会散发大量的热量. 众所周知, 过高的环境温度会损害 IT 设备的运行效率和寿命, 甚至造成 IT 设备损坏. 为了维持 IT 设备的正常运行, 数据中心内需要合理部署制冷系统以吸收由各种用电设备散发的热量. 制冷系统的能耗与 IT 设备的产热量和分布息息相关, 而 IT 设备的产热行为通常由其负载特性决定, 因此, 合理调度 IT 负载对提高制冷系统能效具有重要意义. 暖通空调系统 (heating, ventilation, and air conditioning, HVAC) 和机房空调系统 (computer room air conditioner, CRAC) 是两种常用的数据中心制冷系统, 用于监测和维护数据中心的温度、气流分布和湿度. 本小节主要对基于 IT 负载调度的数据中心空调系统功耗管理途径进行总结和分析, 包括机柜间热量均衡、制冷代价优化、峰值热负载削减和利用免费冷源 4 个方面.

3.1.1 机柜间热量均衡

数据中心广泛采用架空地板的空调制冷方式. 具体而言, 机架上的风扇将冷气从两排机架中间地板下方的小孔抽到机架内来冷却正在工作的服务器, 而被转移的热量将被释放到机架另一侧, 并被顶部排气系统抽走以用于制冷循环, 形成冷热通道相间部署模式. 热气会在相邻的两排机柜顶部的中间位置形成热点 (hot spot), 而热点的温度则取决于相关服务器的设计工艺以及对应的负载水平. 在采用气冷制冷的数据中心内, 热点问题不可避免. 服务器的正常运行需要维持在一定的温度范围内, 为了保证热点区域内的服务区最高温度不超过允许值, 制冷系统必须消耗更多能耗以提供足够的制冷量. 因此, 局部服务器温度过高会降低制冷系统的制冷效率, 从而造成更多的制冷系统能耗浪费^[13]. Sharma 等^[17]通过实验验证了不均匀的机柜间温度分布将导致制冷效率低下, 并提出一种负载调度方法可以促进热区域内温度的均匀分布. 综上所述, 均衡机柜间的热量分布、降低机柜入口温度成为降低制冷需求、提高制冷系统能效的重要方法之一.

再循环 (recirculation) 是影响数据中心机柜间热量均衡的一个重要因素. 当冷通道提供的冷量不足时, 相邻热通道的部分热气会涌入到冷通道进行再循环, 导致冷通道内的冷量进一步减少, 严重降低制冷系统的能效. Moore 等^[18]首次探索利用温度感知的负载调度方法来消除再循环对制冷能效的影响. 他们提出两种基于热力学公式、任务功率画像 (job power profiles) 以及机柜入口出口温度的负载调度策略 ZBD 和 MinRH 以最小化最大的机柜入口温度. ZBD 的基本思想是服务器负载水平对其温度有直接影响关系, 因此, 负载的放置量应该与服务器的入口温度成反比例; 而 MinRH 的基本思想则是机柜当前温度不仅与其负载有关, 还与附近发生的再循环有关, 因此, 他们提出了一种指标 HRF (heat recirculation factor) 来衡量再循环对机柜的程度, 并尽量将负载放置在受再循环影响最小的机柜. 实验表明, 上述两种策略能够使数据中心制冷成本减半. 然而, 上述方法存在一些不足, 首先, 为数据中心内类型众多的任务进行功率画像很难实现^[19]; 其次, 上述方法中采用的热量和功率模型在数据中心的场景下并不精确^[20].

目前, 描述数据中心热量分布最精确的方式是在给定服务器的布局、对应的 CPU 利用率等信息的前提下, 使用计算流体动力学 (computational fluid dynamics, CFD) 建模的方法预测出数据中心的温度分布、绘制出对应的热量分布图 (thermal map)^[21~23]. Bash 等^[23]考虑负载的空间分布对制冷效率的影响问题, 用 CFD 模型预测出惠普公司的一个数据中心的热力分布情况, 并得出数据中心内的某些物理位置比其他物理位置的冷却效率更高的结论. 因此, 他们认为, 将大量的计算工作负载分配到更高效的服务器上的做法能够带来更大的制冷效率提升. 在该想法下, 他们提出一种基于制冷效率的负载调度策略. 该策略首先衡量特定位置上的服务器的制冷效率并以此对所有服务器进行排序, 然后按顺序调度任务, 并且把处于空闲状态的服务器关闭. 他们在一个真实数据中心的一个热力隔离区域进行了一系列的实验, 来测试上述方法的节能潜力, 实验结果表明, 该方法可以为数据中心节省 32% 的总能耗.

遗憾的是, 基于 CFD 构建热量分布图的方法的计算开销太大, 无法指导在线的任务调度^[20, 24]. 围绕这问题, 人们又提出一系列的探索^[19, 20, 25~31].

Tang 等^[20]对计算设备的功耗进行了分析和建模, 验证了在同质 (homogeneous) 数据中心环境下, 服务器的 CPU 利用率与其功耗之间存在简单的线性关系. 在此基础上, 他们提出了一种低复杂度的计算方法, 可以由机柜内的服务器利用率向量计算出同质数据中心内任意机柜的入口温度 (inlet temperature). 与基于 CFD 建模的方法相比, 该方法复杂性更低, 可以满足在线调度对预测速度的要求, 且精度在可接受范围内. 利用上述计算方法, 他们将最小化数据中心制冷需求的问题抽象为通过

在服务器间合理调度到达的任务以最小化数据中心内最大机柜入口温度的问题,并相应提出了基于遗传算法的任务分配算法 XInt-GA 和基于序列二次规划的任务分配算法 XInt-SQP 来求解该优化问题. 在小规模的数据中心内的实验结果显示,与其他方法相比,上述方法可以将最大机柜入口温度降低 $2^{\circ}\text{C}\sim 5^{\circ}\text{C}$,节省 20%~30% 的数据中心制冷系统能耗.

Chen 等^[25]也研究了空间感知的任务调度问题,他们提出一种基于模型的强化学习方法来学习和预测在负载的不同分布下对应的热量分布情况,并且利用在真实数据中心内收集的数据对模型进行训练.在此基础上提出一个两阶段的空间感知的负载调度方法,该方法首先依据制冷效率差异将负载分配到不同的机柜内,然后根据机柜内气流的制冷效果将负载分配到机柜内的服务器上.与非空间感知的负载调度方案相比,该方案可以将最大机柜入口温度降低 $2^{\circ}\text{C}\sim 3^{\circ}\text{C}$.

Li 等^[30]提出一种系统的方法来整合数据中心物理规律和传感器观测数据,并设计热量预测模型 ThermoCast,该模型可以根据连续监测的温度和气流数据较为精确地预测出服务器周围的温度.与 CFD 模型相比,ThermoCast 能够在保证相当的预测准确性的同时,降低模型的复杂性.在商业数据中心内的实验结果显示,ThermoCast 比普通的数据驱动的机器学习方法精度高 1 倍,而且能够提前 4.2 min 成功预报热警报.

Wang 等^[19]提出了一种分析模型来描述数据中心资源的热传递特性以及工作负载的散热特性,该分析模型比 CFD 模型复杂度低,可以适用于在线调度,并且可以达到与文献 [20] 中的方法类似的刻画精度.利用该分析模型,他们开发了一种新的热量感知工作负载调度思路:首先预测在各种可能的负载分布下数据中心温度的增量,然后挑选出最佳的负载分布方案,最后更新热量分布信息.在该思路下,他们提出两种热量感知的任务调度算法 TASA 和 TASA-B 以降低最大机柜入口温度.TASA 根据负载的温度画像来调度负载,从而实现降低数据中心温度的目的,TASA-B 则是在 TASA 的基础上进行了扩展,允许在热量不增加的前提下进行任务回填 (backfilling) 操作.仿真结果表明,在不引入明显系统性能损失的前提下,上述两种算法可以显著降低数据中心的温度.

Zhou 等^[26]系统地研究了 Hadoop 集群的再循环问题.他们首先提出了一个 Hadoop 的热量模型,该模型可以根据当前节点 CPU/磁盘的利用率、运行的任务数量,以及运行时所有从属 (slave) 节点给出的反馈信息预测出数据中心的热流行为.在此基础上,他们针对 Hadoop 集群提出了一种热量感知的任务调度器 TAH,TAH 利用上述热量模型提供的反馈信息在全数据中心范围内调度任务以最小化所有节点上最大的入口温度,从而减少数据中心的电力消耗和冷却成本.实验表明,与默认 FIFO 调度器相比,TAH 平均可以使机柜入口温度降低 2.5°C .

3.1.2 制冷代价优化

影响制冷系统能效的因素众多,除单独优化机柜间温度分布外,部分研究者也考虑将影响制冷系统能效的因素整合在一起构建制冷代价优化模型^[15,18,20,24,25,32],通过负载调度的方式从整体上提升数据中心制冷系统能效.Yeo 等^[33]对现代数据中心广泛使用的 HVAC 系统的能耗组成进行了剖析,发现主要的耗能部件是冷却装置 (chiller) 以及驱动冷气为 CPU 等热源散热的风扇 (fan),且二者的功耗占总 HVAC 系统的总功耗比均为 39%.因此,目前的节能研究主要针对这两个节能部件进行.依据节能对象的不同可以将基于 IT 负载调度的制冷系统能效优化研究分为两种:一种只考虑冷却装置的代价最小化^[18,20,32],另一种同时考虑冷却装置和风扇的总体代价最小化^[24,25].

制冷循环的效率可以用 COP (coefficient of performance) 来表示^[18].通常,冷却装置的制冷代价与 COP 密不可分.COP 定义为被冷却装置转移走的热量 Q 和为转移这部分热量而消耗的能量 W 之间的比值.显然,COP 值越大表示冷却装置的制冷效率越高.当服务器能耗固定时,尽量提高冷气供

应温度可以降低制冷系统能耗, 从而提升 COP; 此外, 当服务器能耗与冷气供应温度均固定时, 合理调度机柜间的负载分布避免出现热点也可以减少制冷系统的能耗浪费, 从而提升 COP. Moore 等^[18]通过对 HP 公司的数据中心内的冷水机组进行观察和研究, 发现其制冷效率 COP 可以被形式化为冷气供应温度 T_s 的二项式函数 (即 $\text{COP}_{T_s} = 0.0068T_s^2 + 0.0008T_s + 0.458$). Banerjee 等^[32]考虑了冷却装置的能耗最小化问题. 他们将冷却装置的能耗形式化为一个与空调系统供应气流温度和 COP 系数相关的函数, 并提出一种制冷效率感知的任务调度算法 HTS 来最小化冷却装置的能耗.

Moore 等^[24]考虑了冷却装置和风扇的总体制冷代价. 他们将数据中心的制冷代价形式化为一个与空调系统供应气流温度、COP 系数、风扇转速有关的函数. 借助于机器学习的方法, 他们首先构建了一个可以根据服务器资源利用率、机房空调设置, 以及环境参数自动在线预测数据中心热量分布图的模型 Weatherman. 实验证明, 超过 90% 的情况下 Weatherman 的预测误差可限制在 0.87°C 以内. 在 Weatherman 的基础上, 他们提出一种简单的热量感知任务调度算法来最小化制冷代价. 实验结果表明, 该算法比 MinHR 算法^[18]可多节省 13%~25% 的制冷代价.

3.1.3 峰值热负载削减

随着数据中心的计算密度逐渐增大, 投入在制冷系统中的开支也越来越大. 为了防止出现较高的服务器故障率, 数据中心制冷系统的建设必须能够满足最坏情况下 (例如, 峰值热负载与外界环境温度同时达到非常高的值) 的制冷需求, 因此, 数据中心内的制冷系统资源通常处于过度供应的状态. Manousakis 等^[34]提出将超额订购的思想引入数据中心制冷系统以提高制冷系统的资源利用率. 当超额订购冷量时, 可能会出现制冷系统无法满足峰值热负载的冷却需求的情况. 此时, 通过强行降低 IT 设备处理能力的方式削减峰值热负载会造成服务性能损失; 而放任 IT 设备温度升高则会影响到其可靠性. 为了确定理想的超额订购量, Manousakis 等^[34]提出了一种优化和仿真框架 CoolProvision, 可以在给定性能约束的前提下求出需要提供的最少冷量.

从上述研究中可以看出, 削减热负载峰值对进一步降低数据中心制冷需求、缩减制冷系统规模、提高制冷系统能效有着重要的意义. 具体而言, 一方面, 热负载峰值出现的时间占比通常很低, 削减峰值热负载可以在满足制冷需求的前提下运行更多的服务器, 提高制冷系统的资源利用率; 另一方面, 部分服务器的升级换代可能导致数据中心制冷需求峰值提升, 削减峰值热负载可以使现有制冷系统在不扩容的前提下仍然满足新的制冷需求.

目前, 在数据中心内能够削减峰值热负载而不引起较大性能影响的技术主要有两大类, 一类是在冷量供应充足的情况下借助数据中心的辅助塔 (auxiliary tank) 或者备用冷源以冰、冷却水等形式存储冷量^[35, 36], 在需要的时候将冷量释放出来辅助制冷系统制冷^[37~39]; 另一类是在服务器上部署新型相变材料 (phase change materials, PCM), 利用 PCM 融化吸热、凝固放热的特性在热负载高峰期融化削减热负载峰值, 在热负载低峰期凝固以延迟热量的释放过程^[40~44].

Skach 等^[45]最先探索了在数据中心尺度下通过部署相变材料吸收存储热量的方法实现峰值热负载削减. 他们引入了基于 PCM 的热量时间转换 (thermal time shifting, TTF) 概念, 即在有益的情况下通过储存和释放热量来重塑热负载. 他们将石蜡 (一种 PCM) 放置在服务器内以演示单个服务器上的 TTF 过程, 并且用模拟器的方式研究了在集群和数据中心内的 TTF 过程. 研究表明, 基于 PCM 的 TTF 可以削减 12% 的峰值热负载 (相当于同等制冷系统资源可满足 114.6% 的服务器的制冷需求). 然而, TTF 是一种被动系统, 不能处理多种混合负载, 也不能适应动态变化的负载环境. 在他们后续的研究中^[46], 他们进一步提出了一种基于 PCM 的热量感知的负载调度技术 VMT. VMT 在 TTF 的基础上添加了一个可以主动调节负载分布的组件, 允许对数据中心的热量分布做更大力度的调整. 他

们为 VMT 设计了两种负载调度算法, 并且在模拟的服务器集群中进行了 VMT 的扩展研究. 实验结果表明, 在 TTF 失效的情况下, VMT 仍然可以削减高达 12.8% 的峰值热负载 (相当于同等制冷系统资源还可以满足额外 7000 多台服务器的制冷需求).

3.1.4 利用免费冷源

近年来, 将数据中心建设在合适的地理位置, 以利用外部免费冷源 (例如冷空气、冷水) 转移热量从而提升数据中心能效的方案逐渐受到工业界的青睐. 例如, 将数据中心部署在寒冷地区 (例如北欧、窑洞), 鉴于外界环境温度长期较低, 可以通过使用风扇将过滤后的外界冷空气引入数据中心内, 将热空气引出到外界的方式替代冷却装置 (例如水冷机组、压缩机等) 转移热量^[47], 对风扇转速的合理控制可以将数据中心内设备的温度保持在安全值以下^[48]. 由于冷却装置耗能占数据中心制冷系统总能耗的比例高达 39%, 因此使用外界免费冷源替代冷却装置制冷的方式可以极大地提高数据中心的能效^[10, 49~51]. Weerts 等^[52] 研究表明, 引入外界冷空气作为冷源对传统机房空调系统进行改造后, 制冷系统的能耗在夏季与冬季分别减少了 70% 和 90%. Endo 等^[53] 在东京郊区建立了一个利用外界冷空气与传统冷却装置混合制冷的数据中心, 经过一年时间的运行数据表明, 该数据中心的 PUE 低至 1.058, 与未利用免费冷源的数据中心相比, 每年可节省 20.8% 的能耗. Zhou 等^[54] 也得出了类似的结论, 通过构建真实的混合制冷数据中心, 他们估计每年可以给数据中心带来高达 30% 的能耗节省.

除了外部气候条件外, 内部 IT 负载的分布和变化特性也是限制免费冷源在数据中心内应用的重要因素. 合理调度 IT 负载对提高利用免费冷源的数据中心的制冷系统能效具有非常重要的意义. Kim 等^[55] 提出一种动态功耗管理方案, 该方案利用时变的外部气候条件以及数据中心内部的负载特性, 通过虚拟机合并的技术实现在满足服务水平协议 (service-level agreement, SLA) 需求的前提下最小化整个数据中心能耗的目的. 该方案只在外界冷源温度低于预设阈值的情况下使用外界免费冷源制冷. 实验表明, 上述方案比传统制冷方案可节省 25.7% 的能耗. 此外, 文献^[56, 57] 也验证了利用外界冷空气的昼夜周期变化特性将负载调度与制冷系统管理相结合的方式可以提高数据中心能效. Yeo 等^[58] 提出一种外界环境感知的负载调度方案, 可以在极端环境下实现服务器温度封顶, 从而促进免费冷源在中小型数据中心的安全应用.

使用免费冷源虽然可以给数据中心带来可观的能耗节省, 但是, 该方法也容易影响到数据中心内硬件设备的可靠性. 研究表明, 服务器的绝对温度达到 50°C 就将影响到磁盘的可靠性^[59], 此外, 长时间的温度变化也会显著增加设备的故障率^[60]. 而 Goiri 等^[61] 在研究了一个利用免费冷源的真实数据中心内部绝对温度与温度变化情况之后指出, 单纯节省数据中心能耗的冷量供应方式 (例如文献^[56, 62~64] 中的方法) 会导致局部服务器温度过高或者温度变化频繁, 从而影响硬件的可靠性. 他们提出一种负载和制冷系统联合管理的系统 Coolair, 该系统可以消除免费冷源的利用给硬件稳定性带来的潜在隐患, 从而进一步促进免费冷源在构建绿色数据中心中的应用.

3.1.5 小结

本小节介绍了“基于 IT 负载调度的制冷系统能效优化”的相关研究, 表 2^[18~20, 24~26, 32, 45, 46, 55, 61, 65] 选取了部分典型研究进行了总结. 从该表可以看出, 目前基于 IT 负载调度的制冷系统能效优化主要还是从均匀机柜间热量分布的角度着手. 均匀机柜间热量分布一方面可以避免出现热点, 另一方面留出更大的空间容纳未来的热负载. 削减峰值热负载也可以避免热点, 该方法主要用于在需要进行制冷能力超额订购的场景下保障设备不会过热. 这两种方法都是以温度和热量作为优化目标间接减少制冷系统能耗, 而制冷代价最小化则是直接将制冷系统能耗作为优化目标进行优化. 上述方法都是从提高

表 2 基于 IT 负载调度的制冷系统能效优化的方法对比

Table 2 Comparison of energy efficiency optimization methods of refrigeration system based on IT load scheduling						
Classification	Source	Objective	Constraint	Method	Platform	Result
Heat balance	[18]	Minimize max cabinet inlet temperature	Server capacity	Temperature aware, heuristic	Simulation	Saving 25% cooling cost
	[20]	Minimize max cabinet inlet temperature	Server capacity	Temperature aware, genetic algorithm, sequential quadratic programming	Small data centers	Reduce max cabinet inlet temperature by 2°C~5°C, saving 20%~30% cooling cost
	[19]	Minimize max cabinet inlet temperature	Server capacity, delay	Temperature aware, heuristic	Simulation	Delay increase 11%, reduce average and max temperature by 14.6 F and 4.9 F
	[25]	Minimize max cabinet inlet temperature	Server capacity	Cooling efficiency aware, reinforcement learning	Simulation	Reduce max cabinet inlet temperature by 2°C~3°C
	[26]	Minimize max cabinet inlet temperature	Server capacity	Heat aware, heuristic	Simulation	Reduce max cabinet inlet temperature and cooling cost by 2.5°C and 15%
Reduce cooling cost	[32]	Minimize chiller plant cost	Server capacity, SLA	Air supply temperature	Simulation	Reduce cooling cost by 15%, increase throughput per energy by 6.89%
	[24]	Minimize the cost of chiller plant and fans	Server capacity	Air supply temperature, fan speed	Simulation	Reduce cooling cost by 13%~25%
Peak heat shaving	[45]	Minimize the peak heat	Server capacity, QoS	PCM-TTF	Single server simulation	Reduce peak heat by 12%
	[46]	Minimize the peak heat	Server capacity, QoS	PCM-VMT	Simulation	Reduce peak heat by 12.8% even when TTF fails
Use free cooling	[55]	Minimize cooling cost	Server capacity, SLA	Use free cooling, rolling time domain control	CloudSim simulation [65]	Reduce energy cost by 25.7%
	[61]	Minimize internal temperature variance	Server capacity	Internal temperature variance aware	Simulation	Reduce cooling cost and internal temperature variance

制冷效率、减少冷却需求两个方面来减少制冷系统能耗, 而利用免费冷源则是减少因为消耗电能而产生的冷却系统能耗. 随着数据中心能耗需求进一步增大, 使用免费冷源为数据中心制冷将成为未来的一大趋势.

3.2 基于 IT 负载调度的供电系统能效优化

数据中心的供电系统为数据中心内部所有 IT 设备的正常运行提供动力,合理调度 IT 设备负载在数据中心内的分布对保障用电安全、提高供电系统能效具有重要意义.目前,基于 IT 负载调度的供电系统功耗管理技术主要包括功率限制、UPS 电源补充供电,以及供电设备调度,本节将围绕这 3 种技术对供电系统能效优化领域的最新研究成果进行总结和分析.

3.2.1 功率限制

数据中心供电系统的构建需要花费巨大的财力和人力,为了使一次性设施的成本被更好地均摊,人们希望能够最大限度地利用数据中心设施资源.为了便于功率管理以及保障安全,数据中心内的供电系统广泛采用断路器 (circuit breaker) 对子电路的功率进行限制,形成由上往下的功率分发层次结构^[66].这种功率分发层次结构往往会使运营商过高估计数据中心的电力资源需求量而过低估计数据中心所能容纳的 IT 设备数目.一方面,由于子电路耗电行为的不确定性,数据中心运营商必须保守地为上层电路预定足够多的功率以保证其下每个分支电路能够分配到的功率不低于其最大额定功率;另一方面,子电路内所能容纳的 IT 设备的总数必须足够保证设备总的最大额定功率不超过子电路的最大功率限制.然而,现实中数据中心内 IT 设备的实际运行最大功率普遍低于其最大额定功率,而且,所有服务器在同一时刻运行在其峰值功率的情况很少^[67].因此,数据中心预定的功率额度通常远远大于实际需求,而实际 IT 负载水平却很低,这与最大限度利用数据中心资源的目标相悖.

为了减少数据中心电力资源浪费,提升数据中心能效,人们将超额订购 (over-subscription)^[66,68]的思想引入数据中心供电系统功率管理中.通过放松对分支电路的最大额定功率保障约束,超额订购能够大幅提高数据中心电力资源的利用率.然而,这种模式也存在潜在的风险.例如,当大多数分支电路负载水平较高,电路的总功率超过订购值时,断路器将被触发以切断并隔离过载电路.由于备用柴油发电机的启动需要花费若干分钟,该行为会造成过载电路内所有服务器短暂性断电,给数据中心带来无法估量的损失.

上述问题可以通过功率限制技术 (power capping) 避免,具体做法是:首先给予电路内每一台服务器设置功率上限值 (caps),需要注意的是,子电路内所有服务器的功率上限总和不能超过该子电路的最大功率限制值;然后利用现代服务器通常附带的功率管理功能 (例如 RAPL^[69] 或者 DVFS^[70]) 动态控制每一台服务器的功率不超过对应的功率上限.目前,IBM^[71] 和 HP^[72] 的商用系统已经实现在毫秒的时间尺度上以瓦特的粒度对服务器进行功率控制.文献^[66,73]先后用真实数据评估了功率限制在数据中心电力资源利用率提升以及能耗节省方面的潜力.不同的是,前者关注的是小集群尺度 (例如,机架级别) 而后者关注的是大规模企业级数据中心尺度.文献^[73]通过对正在运营的 9 个企业数据中心内总共 132 台服务器的真实负载数据进行实验,结果显示,采用功率限制后电力预算大幅减少,其中服务器组件功率预算可减少 50%,整个系统功率预算可减少 20%,而由此造成的性能损失绝大多数情况下可忽略不计.文献^[66]收集并分析了 Google 正在运营的数据中心内总共 15000 台服务器为期 6 个月的功率特性数据,结果表明,在整个数据中心范围内,实际能够达到的最大总功率仅仅为理论最大峰值的 40%.功率限制技术能够让管理员安全地利用上述功率缺口,通过超额订购节省电力开销.实验结果表明,采用功率限制技术至少可以为数据中心节省 23% 的耗电量,此外,如果数据中心采购绿色节能服务器使得系统能耗能够与其负载水平相一致的话,在全数据中心范围内采用功率限制技术可以使峰值功率削减至少 30%,能耗减半.

数据中心采用层次化的功率分发结构,每个功率分配单元 (power distribution unit, PDU) 均负责

为一定区域内 (例如一个或若干个机架) 的服务器提供电力, 并且独立地限制本地峰值功率在预设值之下以保证整体用电安全. 这种电力分配模式不可避免会影响功率控制对电力资源利用率的提升效果. 原因有两方面: (1) 当 PDU 管辖范围内服务器总功率持续增加到接近预设功率限制值时, 不论数据中心内其他地方的电力供应是否足以支撑新增加的负载, PDU 都将启用功率限制技术迫使辖区内服务器的总功率不再增长; (2) 在电力输送路径上保持容错能力的必要性也会放大利用率飙升的影响. 因此, 即使数据中心总体的利用率很低, 而且在其他地方也有足够的电力传输能力, 局部的用电峰值都会导致该区域内发生持续的功率限制. Pelley 等^[74]认为, 导致这一现象的根源在于数据中心内部各个 PDU 的供电能力仅限于本地电路使用, 不能在全数据中心范围内共享调度. 如果全数据中心范围内所有服务器均能够从任意一台 PDU 上获取电力, 那么数据中心的电力资源可以得到充分利用. 因此, 他们提出一种强连接的重组 (shuffled) 功率分发拓扑, 该拓扑将每台服务器连接到多个 PDU 上, 一方面提高了电力资源分配的灵活性, 另一方面减少了对单个 PDU 故障的电力预留容量要求. 在此功率分发拓扑基础上, 他们设计了基于功率的调度算法, 该算法可以在多个 PDU 之间动态调度 IT 负载, 以充分利用数据中心内的电力资源. 在包含 1000 台服务器的真实数据中心内测试的结果表明, 在几乎不影响服务性能的前提下, 该方案可以减少 47% 的功率预定量, 并且节省数据中心 32% 的建设成本.

上述研究结果表明, 在仿真实验以及小规模数据中心的场景下, 功率限制可以有效提高电力资源利用率、节省耗电量, 以及保障数据中心用电安全. 那么, 在真实的大规模数据中心内应用功率限制技术是否依然有如此效果呢? 2016 年, Facebook 在计算机体系结构顶级会议 ISCA 上发表了由他们开发的功耗管理系统 Dynamo, 已经在 Facebook 所有数据中心内运行了 3 年^[75], 这是向学术界公开的第一个在真实的大规模数据中心中成功应用的功耗监控方案. 该系统解决了若干以往基于仿真实验的方案中未能解决的关键问题, 如: (1) 控制器和被控制端之间通信的可扩展性; (2) 应用端和服务端的功率限制计算; (3) 在异构的工作负载和数据依赖关系下, 多个控制器实例之间的协调问题. Dynamo 是一种由上至下的逐层功率控制系统, 由代理 (agent)、底层子控制器 (leaf controller) 和高层控制器 (higher-level controller) 组成. 其中, 轻量级的代理部署在每一台服务器上, 负责读取该服务器的功率数据发送给对应的底层控制器, 并执行来自底层控制器的功率限制指令. Dynamo 的控制器采用层次化的设计, 对应数据中心的功率分发层次结构. 底层控制器通常对应机架级别 (负责约几百台服务器), 在每个预设的采样周期 (采样周期须大于 2 s 小于 30 s) 内收集来自所有代理的功率数据, 并根据高层控制器下发的功率限制值做出功率限制指令下发给代理. 类似地, 高层控制器从下级控制器收集功率信息, 依据上层控制器下发的功率限制值做出功率限制指令下发给下级控制器. 在 Facebook 包含上万台服务器的大规模商业数据中心内的运行数据显示, 在过去的 6 个月里, Dynamo 已经阻止了 18 次因为无法预料的电力激增而导致的潜在停电事故; 为 Hadoop 集群和搜索集群分别带来了 13% 和 40% 的性能提升; 使数据中心电力资源利用率提升了 8%. 由此可见, 在真实的大规模数据中心场景下, 功率封顶技术依然有着不俗的表现.

为了跨服务器协调功率上限, 上述功率控制决策需要逐层聚合服务器的负载、功率等信息. 从理论上分析, 这类层次化的功率控制方案在大规模数据中心内应用时容易出现响应速度慢、启动延迟大的问题, 不适用于在快速时间尺度上跟踪动态功率限制. 2017 年, Azimi 等^[76]提出一种快速的分散式 (decentralized) 功率限制方案 DPC. 在 DPC 中, 每台服务器上部署了一个代理 (agent), 可以在本地计算当前服务器的功率限制值 (caps) 以使得 (1) 整个集群的聚合吞吐量最大化的同时 (2) 将负载的优先级考虑进去, 并且能够将 (3) 集群的总功率限制在一定的阈值之内. 与已有的功率限制方案相比, DPC 具有两个优势: (1) 采用分散的功率控制框架, 避免集中式控制方案中大规模信息数据流的聚合产生的通信延迟、通信热点以及集中控制器的计算热点等问题, 可显著提高响应速度; (2) 采用分布式

最优化技术进行功率计算,避免简单启发式方案中优化效果差的问题,本地做出的决策可实现全局最优.在由 16 台服务器构成的小集群中的实验结果显示,与 Dynamo 相比,DPC 能够减少 72% 到 86% 的启动延迟(具体值与集群大小有关),提升 16% 的系统吞吐量.遗憾的是文中用来测试的集群规模略小,无法验证该方案在大规模数据中心内的性能表现.

多租户数据中心 (multi-tenant data centers) 是一种常见的数据中心类型,通过为多个租户提供共享的物理设施和资源,多租户数据中心为各行各业提供了一个具有成本效益和可扩展的数据中心解决方案.然而,上述功率限制的方案并不适合多租户数据中心.这是因为在多租户数据中心内,数据中心所有者仅提供物理设备和电力等资源,并不具备服务器和负载的控制权利,而所有服务器的控制权归各个租户所有.对于单个租户而言,在没有激励的情况下要冒着损害服务性能的风险限制租用服务器的峰值功率以成全数据中心总体的电力资源利用率和能效提升显然是不可能的.因此,多租户数据中心的所有者无法利用现有的功率限制方案通过对服务器施加功率限制以超额利用电力资源.显而易见,问题的关键在于同一个数据中心内的多个不同租户之间缺乏合作削减功率峰值的动机.为了解决这个问题,Islam 等^[77]提出了一种基于供给函数招标 (supply function bidding) 的新型市场机制 COOP,该机制利用经济激励的手段刺激和促使不同租户在紧急用电情况下通过协作的方式实现最小化性能损失的前提下满足多个功率限制的约束.理论和实验证明,COOP 是一种双赢的方案,一方面,它通过超额订购增加了数据中心所有者的利润;另一方面,通过在紧急用电情况下对租户控制服务器功率的行为作出经济补偿,降低了租户的成本.遗憾的是,COOP 并不适合在真实多租户数据中心内使用,这是因为 (1) 采用尽力而为 (best-effort) 的设计思想,完全依赖租户自愿参与功率削减来处理紧急用电情况,无法确保功率成功封顶;(2) 由于缺乏对租户服务性能的保障,租户可能会被要求频繁地和/或大量地进行功率削减,这将极大降低租户负载的服务性能.随后,由 Sun 等^[78]提出的一种混合功率供应方案 HyPP 可以解决上述问题.HyPP 允许租户在两种功率提供方式之间进行选择,一种是高可用的功率供应 (high availability),另一种是中等可用的功率供应 (medium availability).前者可以保证百分百的功率供应,而后者需要承担因为超额订购而带来的性能损害甚至是断电的风险.在此模式下,HyPP 解决了当聚合功率超过容量值时 (1) 哪些租户的功率应该被限制? 限制多少? 此外,由于功率限制可能造成性能损失, (2) 对不同租户功率限制时如何保证公平? 以及 (3) 如何在不知道未来电力需求的情况下在线做出封顶决策这 3 个问题.

3.2.2 UPS 电源补充供电

功率限制技术虽然可以在一定程度上提高数据中心电力资源利用率以及能效,但是仍然存在一些问题: (1) 限制服务器的功率会造成服务性能损失,而损失的程度通常与功率受限的程度成正相关;(2) 为了尽量减少由功率限制带来的性能损失,在实际应用中会选取较为保守的限制量(通常仅仅在 2% 到 5% 之间^[66,68,75]),因此不能充分发挥功率限制技术的优势.一种潜在的解决思路是在用电“低谷 (valley)”时期存储电力,在用电“高峰 (peak)”时期利用存储的电力补充服务器因执行功率限制而减少的电力,从而实现消除性能损失的目的.众所周知,电力存储代价非常昂贵,因此,在电网内购置大型蓄电设备以存储电力的方案并无市场.然而,现代数据中心为了应对电网突发的断电和电涌而普遍配备的不间断电源 (uninterrupted power supply, UPS) 可以直接用作上述方案中的电力存储设备以实现无性能损失的功率限制^[79].UPS 是数据中心的重要组成部分,目的是在发生断电或者电涌的时候在几秒钟内成为一个活跃的临时电源替代电网为受影响的物理设施提供不间断的电力供应,直到几分钟后柴油发电机被激活.由于 UPS 电源的蓄电状态关系到数据中心的可用性,此外,频繁的 UPS 电源充/放电行为会加速电源的寿命损耗,因此,在数据中心的可用性与 UPS 电源的寿命之间寻找可

行的权衡是这一类方案需要注意的重要问题。

Govindan 等^[79]首次提出将数据中心内集中部署的 UPS 电源 (即所有用电设备共享同一组 UPS 电源, 通常部署在 PDU 之前) 看作是一个电力缓冲区 eBuff (energy buffer), 在适宜的时候储存电力, 在需要的时候用存储的电力对物理设施用电进行补充。在给定电池的容量/ 生命周期以及数据中心的可用性约束下, 他们通过一系列可行性分析证明了在特定类型的数据中心内利用集中式 UPS 补充供电的方案可行性。具体而言, 该类方案非常适用于功率峰值陡峭且持续时间短的数据中心, 然而, 随着功率峰值持续时间拉长, 受电池容量限制, 这类方案性能会逐渐打折扣, 甚至最后失去效用。此外, 在该方案的配合下, 应用功率限制技术可以在不引入用户性能损失的前提下提高功率限制的力度 (降低 15% 至 45% 的功率预定量, 视峰值功率持续时间长度而定), 进一步提升数据中心电力资源利用率和能效。在该思想下, Wang 等^[80]进一步将电力资源 (包含 UPS 电源) 虚拟化为与计算、网络等一样可供数据中心集中调度的资源, 并提出一个软件管理系统 vPower 来负责在所有应用之间进行电力资源分配、隔离和管理。vPower 允许应用指定自身的功率需求、执行接入控制, 以及部署方案。结合动态监测服务器的功率使用情况, vPower 能够强制调整应用之间的电力分配以满足公平性原则和优化系统效率。实验结果表明, 与传统功率限制方案相比, vPower 能够提升 50% 的系统利用率、提高 12%~28% 的应用性能。此外, vPower 还能用来隔离有电力竞争关系的应用。

除了上述集中式部署外, 分布式部署也是一种常用的 UPS 部署方式。该方式下数据中心内每个/群服务器都配置一组独享的 UPS 电源, 例如, Google²⁾的数据中心为每个服务器均配置了一组 UPS 电源, 此外, Microsoft³⁾和 Facebook⁴⁾为同一个机架配置了一组 UPS 电源。与集中式的 UPS 电源部署方案相比, 分布式的部署方案具有如下几个优势: (1) 不受单点故障问题的影响; (2) 储能容量可以很方便随新服务器的添加而增长; (3) 与集中式部署的 UPS 电源不同, 分布式部署的 UPS 电源通常位于 PDU 之后。由于不需要进行 AC/DC 双向转换, 所以可以避免对应的能量转换损失。上述 3 方面优势使得分布式的 UPS 电源部署方案逐渐受到工业界的青睐。Govindan 等^[81]也研究了利用分布式部署的 UPS 电源补充供电以应对功率限制带来的性能损失问题。通过对电池状态、功率限制以及负载迁移 3 个方面的综合考虑, 他们提出了若干种在线启发式方案, 并且搭建了包含 8 个服务器、8 个分布式部署的 UPS 电源的测试平台对这些方案进行测试。实验结果显示: (1) 分布式部署的 UPS 电源可以补充持续时间较短 (大约 10~20 min 以内) 的功率限制行为; (2) 持续时间超过 30 min 的功率限制需要借助于负载迁移的手段才能不影响服务性能; (3) 此外, 由于仓促的负载迁移本身也存在损害性能的可能性, 因此短时间内的功率限制不建议进行负载迁移; (4) 利用分布式 UPS 电源补充供电结合负载迁移技术的方案可以解决绝大多数情况下由于功率限制带来的性能损失问题。Kontorinis 等^[82]也研究了利用分布式部署的 UPS 电源补充供电解决功率限制带来的性能损失问题。不同的是, 他们研究了更加详细的 UPS 电源补充供电模型, 具体讨论了电池的尺寸、种类, 以及充/放电行为对数据中心可靠性和可用性的影响。在此基础上, 他们提出一种最佳的分布式 UPS 电源补充供电架构, 并且制定对应的延长电池使用寿命的策略。真实的负载数据测试表明, 该方案可以在不引入性能损失的前提下平均削减 19.4% 的服务器峰值功率, 使得数据中心的电力资源利用率提升 24%。

尽管分布式的 UPS 电源部署方案具有诸多优势, 但还是存在一些不足。例如, 每组电池中存储的电能只供固定的一台或者若干台服务器使用, 剩余的储电无法在全数据中心范围内共享, 因此限制了对峰值削减的补充能力。Alanazi 等^[83]提出一种管理方案来解决该问题。该方案是一个两层控制结构,

2) Google server-level UPS for improved efficiency. http://news.cnet.com/8301-1001_3-10209580-92.html.

3) Microsoft reveals its speciality servers, racks, 2011. <http://www.datacenterknowledge.com/archives/2011/04/25/microsoft-reveals-its-speciality-servers-racks/>.

4) Facebook open compute project, 2011. <http://opencompute.org>.

顶层是一个任务调度器,负责决定到达数据中心的虚拟机的部署位置;底层是一个 UPS 控制器,负责控制电池充/放电的时机以及电量.通过在全数据中心范围内合理调度虚拟机的部署位置和 UPS 电源的充/放电行为.该方案可以灵活利用所有的 UPS 电力资源,实现 (1) 在最小化需要处于开启状态的服务器数目的同时,(2) 最小化需要预留的 UPS 电源容量的目的.基于真实负载数据的实验结果表明,跟现有方案相比,该方案每月可以为数据中心节省 20% 的电费.

3.2.3 UPS 能效感知控制

随着机柜/服务器级别的 UPS 电源部署方案在数据中心内的广泛应用,UPS 电源持续消耗的电量也逐渐引起了人们的关注.据 Pakbaznia 等^[12]统计,UPS 电源的总耗电量甚至能达到数据中心总耗电量的 15%.鉴于此,部分研究者^[84~86]开始研究通过 UPS 电源的能效感知控制的方法来提高数据中心供电系统的能效.众所周知,在 UPS 电源内进行 AC/DC 电流转换时会产生一定量的功率损耗.为了减少这部分的功率损耗,美国韦恩州立大学 (Wayne State University) 计算机科学系的施巍松教授团队^[84,85]在他们的数据中心内选取了一个包含 20 台服务器、1 个 UPS 电源的机柜进行观察以研究 UPS 电源的转换效率.基于从该机架收集的数据,他们得出了 UPS 电源转换效率与其负载水平呈正相关的结论,并使用自然对数函数对该关系进行拟合.在此基础上,鉴于 UPS 负载状况取决于其辖区内 IT 设备的负载水平,他们进一步构建出由 IT 负载预测整个机架功率 (包含 UPS 电源的功率损耗) 的机架级功率模型.模拟实验表明,由该模型预测的机架功率与实测功率大致匹配,误差率在 2%~5% 以内.随后,他们构建了一个基于 IT 负载分配的优化模型,通过对该模型的求解,可以最大限度减少 IT 设备的功耗和 UPS 电源的功耗损失.在一个拥有 50 个机架 (1000 个服务器)、运行 10 个应用程序的数据中心内的实验结果表明,与均匀负载分配方案相比,该方案可以节省 5.2% 的功耗.对于一个 PUE 为 1.7, 76 兆瓦级别的数据中心而言,意味着每年可以节省 142.5 万美元的费用.

除在进行电流转换的过程中会造成功率损失之外,影响数据中心内 UPS 电源能效的另外一个重要因素是 UPS 电源的冗余供给.在数据中心内 IT 负载水平较低的时期,IT 负载会被合并到部分服务器上以尽可能多地腾出空闲的服务器,这些处于空闲状态的服务器将会切换到休眠状态以节省能耗.由施巍松等总结的 UPS 电源效率与负载的关系可推出,单纯合并服务器会加剧 UPS 电源的功率损耗,从而带来额外的能耗浪费.AL-Hazemi 等^[86]对目前最新的服务器合并方案进行研究发现,该方案会给数据中心增加额外的 13% 的能耗,而这部分能耗是由 UPS 电源转换功耗损失造成的.为了解决这个问题,他们提出了一种合并非高峰时期使用的 UPS 电源的策略^[86].该策略通过将负载水平低的 UPS 电源的负载全部转移到相邻的 UPS 电源的方式来减少需要开启的 UPS 电源数目,从而间接提高了处于开启状态的 UPS 电源的负载水平 (等同于提高 UPS 电源的转换效率).这是一种自适应的过程,只要相邻的 UPS 电源还有足够的能力容纳转移过来的负载,这个过程就会持续发生.UPS 电源之间的负载转移可以通过在服务器端部署微型自动切换开关 (micro-ATS) 的方式实现.实验证明,该方案可以在服务器合并的节能效果基础上减少 20%~40% 的 UPS 电源转换功耗损耗.因此,综合考虑服务器合并与 UPS 电源合并的节能方案比单独进行服务器合并的方案能节省更多的能耗.此外,该方案还可以减少 20% 的 UPS 电源开启数目.

3.2.4 小结

本小节介绍了“基于 IT 负载调度的供电系统能效优化”的相关研究,表 3^[66,73~77,79~83,85,86]选取了部分典型研究进行了总结.从该表可以看出,与功率限制和 UPS 补充供电相比,UPS 能效感知控制在提高供电系统能耗方面的研究较少,主要原因有两点:(1) UPS 电源耗电量占供电系统总能耗的

表 3 基于 IT 负载调度的供电系统能效优化方法对比

Classification	Source	Objective	Constraint	Method	Platform	Result
Power capping	[73]	Reduce power budget	Server capacity, PDU capacity, QoS	Percentage cut	Enterprise data center	Reduce power budget by 20%, negligible QoS loss
	[66]	Reduce power budget	Server capacity, PDU capacity, QoS	Percentage cut	Google data center	Reduce power consumption 23%
	[75]	Data center power management at scale	Server capacity, PDU capacity	Percentage cut	Facebook data center	Increase QoS and power resource utilization by 13%~40% and 8%
	[74]	Increase power resource utilization	Server capacity, PDU capacity	Reset power distribution topology	Real data center	Reduce power budget and data center construction cost by 47% and 32%
	[76]	Rapid power management	Sever capacity, PDU capacity	Distributed optimization	Small cluster	Reduce setup delay by 72%~86%, increase throughput by 16%
	[77]	Power management of multi-tenant data center	PDU capacity	Supply function bidding	Simulation	Win-win between tenant and owner
UPS supply	[79]	Reduce grid power supply	Battery capacity, lifecycle, availability	Centralized UPS supply	Simulation	Reduce power budget by 15%~45% without user experience degradation
	[81]	Reduce QoS degradation	Battery capacity, server capacity	Distributed UPS supply	Testbed	Can solve performance degradation issue in most sases
	[82]	Reduce QoS loss, extent battery life	Battery capacity, server capacity	Distributed UPS supply	Simulation	Shaving peak power by 19.4%, increase power resource utilization by 24%
	[83]	Increase power resource utilization	Battery capacity, server capacity	Scheduling of load and UPS	Simulation	Reduce electricity cost by 20%
	[80]	Increase power resource utilization	Battery capacity, fairness	Power allocation between applications	Simulation	Increase system utilization and application QoS by 50% and 12%~28%
UPS scheduling	[85]	Reduce the loss of UPS transfer	Battery capacity, server capacity	Loss aware scheduling	Small cluster	Reduce power by 5.2%
	[86]	Reduce the loss of UPS transfer	Battery capacity, server capacity	Loss aware scheduling, UPS sleeping	Simulation	Reduce loss of UPS transfer and amount by 20%~40% and 20%

比值较低 (不超过 15%); (2) UPS 电源的可用性关系到整个供电系统的可用性, 不恰当地调整 UPS 的工作状态可能造成重大故障. 功率限制技术不仅可以提高电力资源利用率, 还可以用于保障数据中心

内设备的用电完全。一般而言,功率限制会对应用性能造成一定的损害,通过合理利用数据中心内广泛存在的 UPS 电源可能实现无性能损失的功率限制,这类方法需要对 UPS 电源容量、电池寿命等多方面进行考量。

3.3 IT 系统与制冷系统跨层统一能效优化

由 3.1 小节不难看出,基于 IT 负载的制冷系统能效优化方案通常会倾向于将 IT 负载分散在所有服务器上以避免数据中心内部出现热点。然而,在所有服务器上分散 IT 负载的思路显然有悖于服务器的负载合并节能原则。因此,单纯针对制冷系统的能效优化方案通常并不能达到提升数据中心总体能效的目的。作为数据中心内两个耗能大户,制冷系统与 IT 系统的耗能行为密切相关。具体而言,IT 设备的负载水平决定了其能耗状况,进而又通过释放的热量与制冷系统的能耗关联。因此,设计合理的 IT 负载调度方案以联合优化 IT 系统与制冷系统的能耗对从整体上提升数据中心能效具有十分重大的意义。为此,需要首先构建出统一的跨层能耗模型。目前,统一的跨层能耗模型主要有两种构建方式:(1) 依据设备的物理能耗特性使用具体的数学公式分别计算 IT 系统与制冷系统的能耗;(2) 利用当前/历史信息(包括负载、空调参数、外界环境参数等)借助学习工具对 IT 系统与制冷系统温度或者能耗进行预测。根据上述能耗模型的构建方式可以将该领域内已有的方案分为两类,即基于解析的跨层联合优化和基于预测的跨层联合优化。本节将按照上述分类对 IT 系统与制冷系统跨层统一能效优化领域的最新研究成果进行总结。

3.3.1 基于解析的跨层联合优化

制冷系统的能耗主要产生自转移 IT 系统散发的热量的过程中,因此,既取决于 IT 负载,又取决于制冷系统的制冷效率 COP。目前,对 IT 系统与制冷系统总体能耗的计算普遍采用的思路是先计算 IT 系统的能耗(具体计算方法详见文献 [7]),再用 IT 系统的能耗与空调的性能系数 COP 的比值计算出对应空调能耗 [7, 12~15, 87~90]。

Pakbaznia 等 [12] 认为,在不考虑服务器和机架合并的情况下单纯提升制冷系统能效的任务调度方案并不一定会减少数据中心的总能耗。他们首先利用上述 IT 系统能耗与空调系统能耗的关系构建了统一的能耗最小化目标,在此基础上将数据中心总能耗最小化的问题形式化为一个整数规划问题,并提出一种启发式算法 MTDP 对该问题进行求解。通过对该问题的求解可以 (1) 为到达的任务选择最佳的服务器;并且 (2) 根据每台服务器上运行的负载状况选择合适的运行电压和频率以最小化服务器能耗;与此同时, (3) 调节空调的冷气供应温度节省空调系统能耗。实验结果证明,跨层联合节能方案 MTDP 比单纯最小化制冷系统能耗的方案 Xint-SQP [20] 可以达到更佳的数据中心能耗节省效果,即使后者将所有处于空闲状态的服务器切换至休眠状态节省能耗,MTDP 算法仍旧可以为数据中心多节省 13% 的能耗。

Ahmad 等 [13] 进一步对现有的数据中心节能方案进行总结和分析后发现,IT 系统节能与制冷系统节能两个目标之间存在权衡。具体而言,针对 IT 系统的节能方案 (spatial subsetting) 往往会使数据中心内负载聚合到尽可能少的服务器上,造成部分服务器温度过高,从而产生更大的制冷需求(在部分情况下,制冷系统的能耗增幅甚至大于服务器的能耗节省量 [87]);而针对制冷系统的节能方案 (inverse-temperature) 又会使数据中心内负载分散在尽可能多的服务器上,导致更多的服务器能耗开销。鉴于此,为了实现降低数据中心总能耗的目的,两个节能目标必须协同优化。他们提出一种数据中心联合节能方案 PowerTrade,该方案开启的服务器数目介于基于 spatial subsetting 的方案以及基于 inverse-temperature 的方案之间,因此可以实现在服务器节能与制冷系统节能之间达到较为理想的权

衡. 具体而言, PowerTrade 按照温度高低将数据中心内的服务器分为冷区域、温区域以及热区域 3 种. 到达的负载会被优先调度到冷区域与温区域内的服务器上, 只有当上述服务器的空间不够用时, 才会开启热区域内的服务器. 冷、温区域之间的负载分配则按照校准测试 (calibration run) 结果按比例进行. 冷区域内由于冷却效率高, 负载被尽量合并以减少需要开启的服务器数目; 而温区域内为了避免出现热点, 负载被均匀分散在所有的服务器上. 基于真实数据中心负载的实验结果显示, 与单纯节能服务器能耗的方案以及单纯节省制冷系统能耗的方案相比, PowerTrade 可以在引入响应时间开销不超过 1.7% 的前提下实现 30% 的数据中心能耗节省.

除了休眠和电压频率缩放之外, P-states (performance states) 也是一种常用的处理器节能技术. 不同的 P-states 允许处理器在不同的处理速率和能耗之间进行权衡. Al-Qawasmeh 等^[14] 将处理器的 P-states 整合进了数据中心能耗模型中, 并且针对服务性能 (以在截止时间之前完成处理的任务数目衡量) 和数据中心总能耗两种优化目标分别设计了对应的优化方案. 通过合理调度 (1) 处理器的 P-states, (2) 分配给处理器的任务量, 以及 (3) 空调冷气供应温度, 两种方案可以在数据中心范围内实现性能与能耗的协同优化. 实验数据显示, 与文献 [91] 中的数据中心整体优化方案 (仅考虑休眠, 未考虑 P-states 分配) 相比, 该方案可实现平均 17% 的性能提升, 平均 9% 的数据中心总能耗节省.

Wan 等^[15] 研究了跨越芯片层 (chip layer)、服务器层 (server layer), 以及机房层 (room layer) 的跨层数据中心总能效优化框架. 他们将最小化数据中心总能耗的问题形式化为一个混合整数非线性规划问题, 并且提出一种高效的算法 JOINT 进行求解. JOINT 利用跨不同控制层的系统组件之间的协作减少数据中心总能耗, 并能够根据当前工作负载动态配置系统参数. 具体而言, 在芯片层, JOINT 通过 DVFS 技术调节 CPU 的电压和频率, 以实现在维持可接受的 QoS 性能下提升能耗的目的; 在服务器层, JOINT 通过负载的分配使服务器能够在休眠与工作状态之间动态切换, 此外, JOINT 还可以动态改变机架上风扇的转速, 以使内部服务器获得理想的温度状态; 在机房层, JOINT 通过控制冷气供应温度实现对制冷系统能耗的控制. 基于真实流量数据的实验结果显示, JOINT 可以为数据中心节省 4% 的能耗.

根据 Abts 等^[92] 的推测, 未来数据中心网络能耗与总能耗的比重将会增大. Zheng 等^[90] 将日益增长的网络能耗也考虑在数据中心总能耗最小化的目标之内. 实际上, 数据中心内的服务器与网络之间存在紧密的关联: 一方面, 无视网络拓扑结构的服务器合并方案可能会导致流量堵塞, 从而损害网络性能; 另一方面, 服务器合并后会改变网络拓扑结构, 也可能带来更大的节能空间 (例如可以使部分网络交换机/路由器切换至休眠状态). 因此, 为了达到更好的节能效果, 服务器应该与网络协同考虑. 鉴于此, 他们提出一种利用负载的相关性降低数据中心服务器与网络总能耗的联合节能框架 PowerNetS, 并将该框架做了进一步扩展以最小化包含服务器、网络, 以及制冷系统在内的数据中心总能耗. 实验结果显示, 与同样利用负载相关性的服务器节能方案^[93] 相比, PowerNetS 可多降低 44.3% 的能耗. 扩展后的 PowerNetS 与扩展前相比, 可进一步降低 8.8%~14.6% 的能耗.

与上述借助 COP 关联 IT 系统与制冷系统能耗的研究不同, 一些研究者从网宇实体 (cyber-physical) 系统的角度探索了对数据中心总体能耗的管理问题^[16,91,94]. Parolini 等^[91] 提出一种面向控制 (control-oriented) 的模型来评估协作的网宇实体系统策略对数据中心的影响. 该模型将数据中心描述为两个耦合的网络: 一个代表网络动力学的计算网络, 另一个代表物理动力学的热力网络. 两个网络通过 IT 负载相互耦合, 具体而言, IT 负载在服务器上的分布既影响计算网络提供的 QoS, 又影响热力网络产生的热量. 在该模型的基础上, 他们提出 3 种控制策略以从整体上降低数据中心总能耗. Li 等^[16] 的研究目标是为数据中心总能耗最小化问题提供一种闭环的 (closed-form) 解析最优解 (analytic optimal solution). 他们首先构建了简化的 IT 系统和制冷系统能耗模型, 在此基础上提出了

一种最优的负载分布方案. 在真实数据中心内的实验证明, 上述简化的系统能耗模型足以促成不小的数据中心总体能耗节省. 通过合理选择开启的服务器集合、调度负载在服务器上的分布以及制冷系统的冷气供应温度, 上述整体节能方案比未从整体上优化数据中心的方案可以多节省 5%~18% 的能耗.

3.3.2 基于预测的跨层联合优化

由于数据中心内各种系统 (例如机械、电气、控制) 之间有很复杂的动态交互关系, 基于公式的模型无法精确地计算出数据中心的总体能耗^[95,96]. 基于 CFD 仿真软件的方法可以实现高精度能耗预测. 但由于计算复杂度太大, 这类方法无法指导大规模数据中心内实时的负载调度. 为了解决这个问题, Rossi 等^[97] 提出使用 Holt-Winters 算法预测数据中心的未来功耗. 与 CFD 模型相比, 该算法计算复杂度低. 实验数据表明, 该算法对一天以内功耗值的预测误差率在 2% 以内. Hsu 等^[98] 则选择了近几年发展迅猛的机器学习方法, 事实上, 数据中心场景的复杂性以及大量存在的监测数据很适合机器学习的应用. 他们提出一种基于自感知计算 (self-aware computing) 的负载预测框架 SAWF 以预测数据中心整体能耗, SAWF 主要包含 3 个部分, 首先是 (1) 特征抽取模块, 用于评估数据中心内所有设备中变量的重要性, 并动态选择最重要的最相关的变量集合作为数据输入; 在此基础上, (2) 提出一种精确且高效的神经网络模型以预测数据中心未来总能耗; 最后, (3) 设计一种在线的误差监控和模型更新模块, 可以连续监测预测误差, 并且在必要的时候更新上述预测模块. 实验结果显示, SAWF 比其他经典的时间序列预测方法更能准确地预测数据中心总功耗. 除此之外, 还有一些单独预测 IT 系统能耗^[95,99,100] 与制冷系统能耗^[101] 的工作, 这些研究也主要采用机器学习的方法, 上述研究证明了机器学习在数据中心系统能耗建模中的有效性.

不同于上述对数据中心总能耗的预测, Gao 等^[96] 选择直接预测数据中心的 PUE. 他们提出一个神经网络框架, 该框架利用实际数据中心的数据进行学习, 从而构建出数据中心性能模型和 PUE 模型. 实验证明, 该框架可以非常精确预测出 PUE 的值 (对于 PUE = 1.1 的情况, 预测误差仅为 0.4%). 该模型已在谷歌的数据中心内进行了广泛的测试和验证. 该研究也进一步证明了机器学习是利用现有传感器数据对数据中心能效进行建模的有效方法.

3.3.3 小结

本小节介绍了“IT 系统与制冷系统跨层统一能效优化”的相关研究, 表 4^[12~16, 20, 90, 91, 96, 102] 选取了部分典型研究进行了总结. 从该表可以看出, 虽然基于解析的方法精度不如基于预测的方法高, 目前基于解析的方法仍然远远多于基于预测的方法. 这主要是因为基于解析的方法计算简单方便, 不需要借助大规模的真实数据集进行训练. 然而, 随着机器学习领域的发展, 基于预测的方法对于大规模数据集的依赖程度将越来越低, 而其精度高、测试速度快的优势将越来越明显. 可以预见, 未来人们对高能效数据中心的不断追求将进一步促进基于预测的跨层统一优化研究的发展.

4 未来展望

随着云计算在全球范围内的蓬勃发展, 数据中心的数目和规模都得到了前所未有的增长, 而其产生的巨大能耗也成为了国内外关注的热点问题. 数据中心的能耗问题亟待解决, 这是因为: (1) 目前数据中心资源利用率普遍极低, 导致大量的能耗被浪费, 这与全社会面临能源枯竭、环境污染问题的现状形成鲜明对比; (2) 巨额的电费开支给数据中心运营商带来了沉重的负担, 已成为数据中心进一步发展的绊脚石; (3) 为贴近用户提高服务质量, 大多数数据中心倾向于建立在经济发达地区, 这将加剧当

表 4 IT 系统与制冷系统跨层统一能效优化的方法对比
Table 4 Comparison of cross layer energy efficiency optimization methods

Classification	Source	Objective	Constraint	Method	Platform	Result
Analytic approach	[12]	Minimize server and cooling energy	Server capacity	DVFS, server sleeping, air supply	MATLAB	Save 13% more energy than [20]
	[13]	Minimize server and cooling energy	Server capacity	Server sleeping, air supply	Air-PAK ^[102]	Save 30% energy with delay less than 1.7%
	[14]	Maximize server QoS,	Server capacity	P-states, server sleeping, air supply	Simulation	Average 17% performance increase, 9% energy reduction
		Minimize server and cooling energy				
	[15]	Minimize server and cooling energy	Server capacity, QoS	DVFS, server sleeping, fan speed, air supply	Simulation	Reduce energy by 4%
	[90]	Minimize server, network and cooling energy	Server capacity, network bandwidth	Server and network sleeping, air supply	Hardware testbet	Reduce energy by 44.3%, more 8.8%~14.6% reduction with network
Prediction approach	[91]	Minimize server and cooling energy	Server capacity	DVFS, server sleeping, fan speed, air supply	Simulation	Reduce energy by 4%
	[16]	Minimize server and cooling energy	Server capacity, max temperature	Server sleeping, fan speed, air supply	Hardware testbed	Reduce energy by 5%~18%
	[96]	Minimize PUE	Resource capacity	19 types of data center features	Google data center	Reduce significant energy

地电网的供电压力, 从而影响新增数据中心的建设, 进一步阻碍了数据中心的发展. 伴随着云计算时代的到来, 作为云计算重要物理依托的数据中心, 可以预计在未来数年内其能效研究还会持续受到学术界和工业界的广泛关注.

国内在数据中心能效研究方面起步虽晚, 但是近年来发展迅速, 目前已基本与国际水平持平. 随着云计算和大数据被纳入国家发展战略, 加强数据中心能效研究、发挥产学研用联合优势形成自主可控的高效能云计算数据中心核心关键技术体系对促进我国云计算发展有着重要的意义.

单一层面的能效研究在提高数据中心整体能效方面的效果有限, 为了进一步发掘数据中心节能的优化空间, 需要从跨层能效优化的角度入手, 兼顾 IT 系统层和配套设施层的特性. 可以预见, 在未来的研究工作中, 数据中心跨层能效优化有如下 3 个方面值得深入探讨:

(1) 基于 IT 负载调度的制冷和供电系统能效优化问题在过去十余年里受到了广泛的研究, 而 IT 系统与配套设施跨层统一能效优化虽然节能潜力更大, 但是相关研究在近几年才兴起, 因此目前已发表的研究工作较少, 预计该领域的研究在未来数年里将会出现更多的创新成果.

(2) 机器学习对于非线性复杂系统具有良好的适应力, 非常适合处理数据中心场景下各种系统复

杂的动态交互. 随着机器学习技术的不断创新和深入, 利用机器学习领域的创新成果解决数据中心的能效优化问题将成为一个趋势.

(3) 数据中心研究领域的研究成果可以在数据中心内进行部署应用, 借助企业数据中心平台推动高效能数据中心核心关键技术体系结构的研发、促进系统解决方案和标准规范的完善也是重要的研究方向.

参考文献

- 1 Delforge P. America's Data Centers Consuming and Wasting Growing Amounts of Energy. NRDC Technical Report. 2014
- 2 Siriwardana J, Jayasekara S, Halgamuge S K. Potential of air-side economizers for data center cooling: a case study for key Australian cities. *Appl Energy*, 2013, 104: 207–219
- 3 Tantar A A, Tantar E. A survey on sustainability in ICT: a computing perspective. In: *Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation*, Vancouver, 2014. 1213–1220
- 4 Zhang W W, Wen Y G, Wong Y W, et al. Towards joint optimization over ICT and cooling systems in data centre: a survey. *IEEE Commun Surv Tut*, 2016, 18: 1596–1616
- 5 Deng W, Liu F M, Jin H, et al. Leveraging renewable energy in cloud computing datacenters: state of the art and future research. *Chin J Comput*, 2013, 36: 582–598
- 6 Wan J X, Gui X, Zhang R, et al. Joint cooling and server control in data centers: a cross-layer framework for holistic energy minimization. *IEEE Syst J*, 2018, 12: 2461–2472
- 7 Dayarathna M, Wen Y G, Fan R. Data center energy consumption modeling: a survey. *IEEE Commun Surv Tut*, 2016, 18: 732–794
- 8 Wikipedia contributors. Electrical efficiency. 2018. https://en.wikipedia.org/wiki/Electrical_efficiency
- 9 Belady C, Rawson A, Pflueger J, et al. Green Grid Data Center Power Efficiency Metrics: PUE and DCIE. Technical Report, Green Grid. 2008
- 10 Google. Efficiency: How We Do It. Google Technical Report. 2018
- 11 Pelley S, Meisner D, Wenisch T F, et al. Understanding and abstracting total data center power. In: *Proceedings of Workshop on Energy-Efficient Design*, 2009
- 12 Pakbaznia E, Pedram M. Minimizing data center cooling and server power costs. In: *Proceedings of ISLPED*, 2009. 145–150
- 13 Ahmad F, Vijaykumar T N. Joint optimization of idle and cooling power in data centers while maintaining response time. In: *Proceedings of ASPLOS*, 2010. 243–256
- 14 Al-Qawasmeh A M, Pasricha S, Maciejewski A A, et al. Power and thermal-aware workload allocation in heterogeneous data centers. *IEEE Trans Comput*, 2015, 64: 477–491
- 15 Wan J X, Gui X, Kasahara S, et al. Air flow measurement and management for improving cooling and energy efficiency in raised-floor data centers: a survey. *IEEE Access*, 2018, 6: 48867–48901
- 16 Li S, Le H K, Pham N, et al. Joint optimization of computing and cooling energy: analytic model and a machine room case study. In: *Proceedings of ICDCS*, 2012. 396–405
- 17 Sharma R K, Bash C E, Patel C D, et al. Balance of power: dynamic thermal management for Internet data centers. *IEEE Internet Comput*, 2005, 9: 42–49
- 18 Moore J D, Chase J S, Ranganathan P, et al. Making scheduling “cool”: temperature-aware workload placement in data centers. In: *Proceedings of USENIX Annual Technical Conference*, 2005. 61–75
- 19 Wang L, Khan S U, Dayal J. Thermal aware workload placement with task-temperature profiles in a data center. *J Supercomput*, 2012, 61: 780–803
- 20 Tang Q H, Gupta S K S, Varsamopoulos G. Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: a cyber-physical approach. *IEEE Trans Parallel Distrib Syst*, 2008, 19: 1458–1472

- 21 Beitelmal A H, Patel C D. Thermo-fluids provisioning of a high performance high density data center. *Distrib Parallel Databases*, 2007, 21: 227–238
- 22 Choi J, Kim Y, Sivasubramaniam A, et al. A CFD-based tool for studying temperature in rack-mounted servers. *IEEE Trans Comput*, 2008, 57: 1129–1142
- 23 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 USENIX Annual Technical Conference*, 2007. 363–368
- 24 Moore J D, Chase J S, Ranganathan P. Weatherman: automated, online and predictive thermal mapping and management for data centers. In: *Proceedings of ICAC*, 2006. 155–164
- 25 Chen H, Kesavan M, Schwan K, et al. Spatially-aware optimization of energy consumption in consolidated data center systems. In: *Proceedings of American Society of Mechanical Engineers Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems*, 2011. 461–470
- 26 Zhou Y, Taneja S, Dudeja G, et al. Towards thermal-aware Hadoop clusters. *Future Generation Comput Syst*, 2018, 88: 40–54
- 27 Sano K, Shimizu H, Kondo Y, et al. Improving accuracy of temperature distribution and energy-saving technology of air conditioners in data centers. *IEEE Trans Compon Packag Manufact Technol*, 2018, 8: 811–817
- 28 Jiang Z G, Huang W, You I, et al. Thermal-aware task placement with dynamic thermal model in an established datacenter. In: *Proceedings of IMIS*, 2014. 1–8
- 29 Cao T, Huang W, He Y, et al. Cooling-aware job scheduling and node allocation for overprovisioned HPC systems. In: *Proceedings of IPDPS*, 2017. 728–737
- 30 Li L, Liang C L, Liu J, et al. ThermoCast: a cyber-physical forecasting model for datacenters. In: *Proceedings of KDD*, 2011. 1370–1378
- 31 Wei T S, Wang Y Z, Zhu Q. Deep reinforcement learning for building HVAC control. In: *Proceedings of DAC*, 2017. 1–6
- 32 Banerjee A, Mukherjee T, Varsamopoulos G, et al. Integrating cooling awareness with thermal aware workload placement for HPC data centers. *Sustain Comput-Inf Syst*, 2011, 1: 134–150
- 33 Yeo S, Lee H H S. Peeling the power onion of data centers. In: *Energy Efficient Thermal Management of Data Centers*. Boston: Springer, 2012. 137–168
- 34 Manousakis I, Goiri I, Sankar S, et al. CoolProvision: underprovisioning datacenter cooling. In: *Proceedings of SoCC*, 2015. 356–367
- 35 Garday D, Housley J. Thermal Storage System Provides Emergency Data Center Cooling. White Paper Intel Information Technology, Intel Corporation. 2007
- 36 Wang Y F, Wang X R, Zhang Y W. Leveraging thermal storage to cut the electricity bill for datacenter cooling. In: *Proceedings of the 4th Workshop on Power-Aware Computing and Systems*, 2011. 8
- 37 Zhang Y W, Wang Y F, Wang X R. TEstore: exploiting thermal and energy storage to cut the electricity bill for datacenter cooling. In: *Proceedings of CNSM*, 2012. 19–27
- 38 Zheng W L, Ma K, Wang X R. Exploiting thermal energy storage to reduce data center capital and operating expenses. In: *Proceedings of HPCA*, 2014. 132–141
- 39 Roth K, Zogg R, Brodrick J. Cool thermal energy storage. *Ashrae J*, 2006, 48: 94–96
- 40 Raghavan A, Luo Y, Chandawalla A, et al. Computational sprinting. In: *Proceedings of HPCA*, 2012. 249–260
- 41 Raghavan A, Luo Y, Chandawalla A, et al. Computational sprinting on a hardware/software testbed. In: *Proceedings of ASPLOS*, 2013. 155–166
- 42 Raghavan A, Emurian L, Shao L, et al. Utilizing dark silicon to save energy with computational sprinting. *IEEE Micro*, 2013, 33: 20–28
- 43 Raghavan A, Luo Y, Chandawalla A, et al. Designing for responsiveness with computational sprinting. *IEEE Micro*, 2013, 33: 8–15
- 44 Lei S, Raghavan A, Emurian L, et al. On-chip phase change heat sinks designed for computational sprinting. In: *Proceedings of Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM)*, San Jose

2014. 29–34
- 45 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 ISCA*, 2015. 439–449
 - 46 Skach M, Arora M, Tullsen D M, et al. Virtual melting temperature: managing server load to minimize cooling overhead with phase change materials. In: *Proceedings of ISCA*, 2018. 15–28
 - 47 Zhang H N, Shao S Q, Xu H B, et al. Free cooling of data centers: a review. *Renew Sustain Energy Rev*, 2014, 35: 171–182
 - 48 Park J. Open Compute Project-Data Center. Facebook Technical Report. 2011
 - 49 Facebook. Prineville - Facebook power. 2013. <https://fbpuewue.com/prineville>
 - 50 Miller R. The evolution of Facebook's data center cooling. 2012. <http://www.datacenterknowledge.com/archives/2012/12/04/evolution-of-facebooks-cooling/>
 - 51 Gebrehiwot B, Aurangabadkar K, Kannan N, et al. CFD analysis of free cooling of modular data centers. In: *Proceedings of SEMI-THERM*, 2012. 108–111
 - 52 Weerts B A, Gallaher D, Weaver R, et al. Green data center cooling: achieving 90% reduction. In: *Proceedings of GTC*, 2012. 1–6
 - 53 Endo H, Kodama H, Fukuda H, et al. Effect of climatic conditions on energy consumption in direct fresh-air container data centers. *Sustain Comput-Inf Syst*, 2015, 6: 17–25
 - 54 Zhou R L, Wang Z K, McReynolds A, et al. Optimization and control of cooling microgrids for data centers. In: *Proceedings of ITherm*, 2012. 338–343
 - 55 Kim J, Ruggiero M, Atienza D. Free cooling-aware dynamic power management for green datacenters. In: *Proceedings of HPCS*, 2012. 140–146
 - 56 Kumari N, Shah A, Bash C, et al. Optimizing data center energy efficiency via ambient-aware IT workload scheduling. In: *Proceedings of ITherm*, 2012. 539–544
 - 57 Zapater M, Turk A, Moya J M, et al. Dynamic workload and cooling management in high-efficiency data centers. In: *Proceedings of IGSC*, 2015. 1–8
 - 58 Yeo S, Hossain M M, Huang J, et al. ATAC: ambient temperature-aware capping for power efficient datacenters. In: *Proceedings of SoCC*, 2014. 1–14
 - 59 Pinheiro E, Weber W D, Barroso L A. Failure trends in a large disk drive population. In: *Proceedings of FAST*, 2007. 17–23
 - 60 El-Sayed N, Stefanovici I A, Amvrosiadis G, et al. Temperature management in data centers. *SIGMETRICS Perform Eval Rev*, 2012, 40: 163
 - 61 Goiri I, Nguyen T D, Bianchini R. CoolAir: temperature- and variation-aware management for free-cooled datacenters. In: *Proceedings of ASPLOS*, 2015. 253–265
 - 62 Arlitt M, Bash C, Blagodurov S, et al. Towards the design and operation of net-zero energy data centers. In: *Proceedings of ITherm*, 2012. 552–561
 - 63 Liu Z H, Chen Y, Bash C, et al. Renewable and cooling aware workload management for sustainable data centers. In: *Proceedings of SIGMETRICS*, 2012. 175–186
 - 64 Mukherjee T, Banerjee A, Varsamopoulos G, et al. Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. *Comput Netw*, 2009, 53: 2888–2904
 - 65 Buyya R, Ranjan R, Calheiros R N. Modeling and simulation of scalable cloud computing environments and the cloudsims toolkit: challenges and opportunities. In: *Proceedings of HPCS*, 2009. 1–11
 - 66 Fan X, Weber W D, Barroso L A. Power provisioning for a warehouse-sized computer. In: *Proceedings of ISCA*, 2007. 13–23
 - 67 Bhattacharya A A, Culler D, Kansal A, et al. The need for speed and stability in data center power capping. *Sustain Comput-Inf Syst*, 2013, 3: 183–193
 - 68 Fu X, Wang X R, Lefurgy C. How much power oversubscription is safe and allowed in data centers. In: *Proceedings of ICAC*, 2011. 21–30

- 69 Kidd T L. Power Management States: P-States, C-States, and Package C-States. Intel Development Resources, 2014
- 70 Lo D, Cheng L, Govindaraju R, et al. Towards energy proportionality for large-scale latency-critical workloads. In: Proceedings of ISCA, 2014. 301–312
- 71 Popa P. Managing Server Energy Consumption Using IBM PowerExecutive. IBM Techique Report, 2006
- 72 Staff H P. HP Power Capping and Dynamic Power Capping for ProLiant Servers. HP Techique Report TC090303TB, 2009
- 73 Ranganathan P, Leech P, Irwin D E, et al. Ensemble-level power management for dense blade servers. In: Proceedings of ISCA, 2006. 66–77
- 74 Pelley S, Meisner D, Zandevakili P, et al. Power routing: dynamic power provisioning in the data center. In: Proceedings of ASPLOS, 2010. 231–242
- 75 Wu Q, Deng Q Y, Ganesh L, et al. Dynamo: Facebook's data center-wide power management system. In: Proceedings of ISCA, 2016. 469–480
- 76 Azimi R, Badiei M, Zhan X, et al. Fast decentralized power capping for server clusters. In: Proceedings of HPCA, 2017. 181–192
- 77 Islam M A, Ren X, Ren S L, et al. A market approach for handling power emergencies in multi-tenant data center. In: Proceedings of HPCA, 2016. 432–443
- 78 Sun Q H, Ren S L, Wu C. Fair online power capping for emergency handling in multi-tenant cloud data centers. IEEE Trans Cloud Comput, 2017. doi: 10.1109/TCC.2017.2762311
- 79 Govindan S, Sivasubramaniam A, Urgaonkar B. Benefits and limitations of tapping into stored energy for datacenters. In: Proceedings of ISCA, 2011. 341–352
- 80 Wang D, Ren C G, Sivasubramaniam C. Virtualizing power distribution in datacenters. In: Proceedings of ISCA, 2013. 595–606
- 81 Govindan S, Wang D, Sivasubramaniam S, et al. Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters. In: Proceedings of ASPLOS, 2012. 75–86
- 82 Kontorinis V, Zhang L E, Aksanli B, et al. Managing distributed UPS energy for effective power capping in data centers. In: Proceedings of ISCA, 2012. 488–499
- 83 Alanazi S, Dabbagh M, Hamdaoui B, et al. Reducing data center energy consumption through peak shaving and locked-in energy avoidance. IEEE Trans Green Commun Netw, 2017, 1: 551–562
- 84 Zhang Q, Shi W S. UPS-aware workload placement in enterprise data centers. IEEE Comput Mag, 2015, 1: 1–7
- 85 Zhang Q, Shi W S. Energy-efficient workload placement in enterprise datacenters. Computer, 2016, 49: 46–52
- 86 AL-Hazemi F, Peng Y, Youn C H, et al. Dynamic allocation of power delivery paths in consolidated data centers based on adaptive UPS switching. Comput Netw, 2018, 144: 254–270
- 87 Abbasi Z, Varsamopoulos G, Gupta S K S. TACOMA: server and workload management in internet data centers considering cooling-computing power trade-off and energy proportionality. ACM Trans Archit Code Opt, 2012, 9: 1–37
- 88 Antal M, Pintea C, Pintea E, et al. Thermal aware workload consolidation in cloud data centers. In: Proceedings of ICCP, 2016. 377–384
- 89 Fu L J, Wan J X, Yang J, et al. Dynamic thermal and IT resource management strategies for data center energy minimization. J Cloud Comp, 2017, 6: 25
- 90 Zheng K Y, Zheng W L, Li L, et al. PowerNetS: coordinating data center network with servers and cooling for power optimization. IEEE Trans Netw Serv Manage, 2017, 14: 661–675
- 91 Parolini L, Sinopoli B, Krogh B H, et al. A cyber-physical systems approach to data center modeling and control for energy efficiency. Proc IEEE, 2012, 100: 254–268
- 92 Abts D, Marty M R, Wells P M, et al. Energy proportional datacenter networks. In: Proceedings of ISCA, 2010. 338–347
- 93 Wang X D, Wang X R, Zheng K Y, et al. Correlation-aware traffic consolidation for power optimization of data center networks. IEEE Trans Parallel Distrib Syst, 2016, 27: 992–1006

- 94 Yao J G, Guan H B, Luo J Y, et al. Adaptive power management through thermal aware workload balancing in internet data centers. *IEEE Trans Parallel Distrib Syst*, 2015, 26: 2400–2409
- 95 Li Y L, Hu H, Wen Y G, et al. Learning-based power prediction for data centre operations via deep neural networks. In: *Proceedings of the 5th International Workshop on Energy Efficient Data Centres*, Waterloo, 2016. 1–10
- 96 Gao J, Jamidar R. *Machine Learning Applications for Data Center Optimization*. Google White Paper. 2014
- 97 Rossi M, Brunelli D. Forecasting data centers power consumption with the Holt-Winters method. In: *Proceedings of 2015 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS)*, 2015. 210–214
- 98 Hsu Y, Matsuda K, Matsuoka M. Self-aware workload forecasting in data center power prediction. In: *Proceedings of CCGrid*, 2018. 321–330
- 99 Liu N, Li Z, Xu Z Y, et al. A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning. In: *Proceedings of ICDCS*, 2017. 372–382
- 100 Berral J L, Goiri I, Nou R, et al. Towards energy-aware scheduling in data centers using machine learning. In: *Proceedings of e-Energy*, 2010. 215–224
- 101 Li Y L, Wen Y G, Guan K, et al. Transforming cooling optimization for green data center via deep reinforcement learning. 2017. ArXiv: 1709.05077
- 102 AirPAK. Computational fluid dynamics software by Ansys Inc. <http://www.ansys.com/products/airpak/default.asp>

Research progress and development trend of cross-layer energy efficiency optimization in data centers

Jiye WANG¹, Biyu ZHOU^{2*}, Wantao LIU² & Songlin HU²

1. *State Grid Corporation of China, Beijing 100031, China;*

2. *Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100089, China*

* Corresponding author. E-mail: zhoubiyu@iie.ac.cn

Abstract Energy consumption has become a main obstacle hindering the further development of data centers. To solve this problem, many data center energy efficiency optimization studies have emerged in academia and industry in recent years. This review introduces current global research on the data center energy saving problem from the perspective of optimizing cross-layer energy efficiency, including the refrigeration and power supply system energy efficiency optimization based on IT load dispatching as well as unified energy efficiency optimization of IT and refrigeration systems, and then projects trends in these areas.

Keywords energy consumption, data center, energy efficient algorithms, resource management, green computing

Jiye WANG was born in 1964. He received his Ph.D. degree from the Tsinghua University, Beijing, in 2007. Currently, he is a senior engineer in the Information and Communication Department of the State Grid Corporation of China. His research interests include information, management of power system, smart grids, and next generation energy systems.

Biyu ZHOU was born in 1990. She received her Ph.D. degree from the Institute of Computer Technology, Chinese Academy of Sciences, Beijing, in 2018. She is currently an assistant professor at the Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include green computing, cloud computing, and machine learning.

Wantao LIU was born in 1980. He received his Ph.D. degree from Beihang University, Beijing, in 2013. He is currently an associate professor at the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include distributed systems, service computing, and parallel computing.

Songlin HU was born in 1973. He received his Ph.D. degree from the Beihang University, Beijing, in 2001. Currently, he is a professor at the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include enterprise-level big data processing, distributed systems, and service computing.