绿色数据中心的热量管理方法研究

李翔，姜晓红，吴朝晖，叶可江

（浙江大学计算机科学与技术学院 杭州 中国 310027）

摘 要 数据中心的高能耗是一个亟待解决的问题。尤其是随着云计算的发展，更多的资源集中到云端。构建绿色数据中心、实现节能减排成为了近年来业界关注的热点。数据中心的能耗主要由计算能耗和制冷能耗两部分组成。数据中心的热量管理主要从减少制冷能耗的角度出发，为实现绿色计算提供了新的思路。本文从绿色数据中心的状态监控、热量建模、热量管理策略以及热量管理评价四个方面综述了近年来数据中心热量管理方面的研究工作。本文提出了绿色数据中心热量管理的总体架构，总结了其分布式监控系统的一般框架；对现有的热量管理方法按面向单节点/面向多节点进行分类，并且从复杂度、灵活度、实施效果等多方面进行了比较，分析了各种方法的优势和局限性。本文提出了数据中心全局能耗评价、制冷系统效率评价、热量及温度评价的分类方法，对现有的评价方法进行总结。最后论文列出了未来需要进一步研究的十个方向。

关键词 绿色计算；绿色数据中心；热量管理；能耗管理；制冷；云计算

Research of Thermal Management Methods for Green Data Centers

LI Xiang, JIANG Xiao-Hong, WU Zhao-Hui, YE Ke-Jiang

（College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China）

Abstract The high energy consumption of data center is a serious problem to be solved. Especially as the development of cloud computing, more resources are centralized to the clouds. Constructing green data centers and achieving power cost and carbon footprint reduction become research hotspots in recent years. Energy consumption of data centers consists of computational energy and cooling energy. Thermal management for data centers mainly target reducing cooling energy and provide new solutions to achieving green computing. This paper presents a review of the recent research work of thermal management from the perspectives of status monitoring, thermal modeling, thermal management policies and thermal management evaluations. We propose the overall architecture of thermal management of green data center, and summarize its general framework of distributed monitoring system. We classify the existing thermal management policies into two classes: single-node case and multiple-node case, compare the complexity, flexibility and effectiveness of the existing policies, and analyze their advantages and limitations. This paper summarizes the existing evaluating approaches proposing a new taxonomy of three categories: global energy consumption evaluation, refrigerating efficiency evaluation, and thermal/temperature evaluation. Finally, ten possible research directions are suggested for future research in this field.

Key words green computing; green data center, thermal management; energy management; refrigeration; cloud computing

---

本课题得到国家“八六三”高技术研究发展计划重大项目基金（2011AA01A207）和国家自然科学基金项目（61272128）资助. 李翔（计算机学会会员号：E200031239G），男，1990年生，博士研究生，主要研究方向为云计算、数据中心节能、建模和仿真. E-mail: lixiang2011@zju.edu.cn. 姜晓红（通信作者），女，1966年生，博士，副教授，主要研究方向为计算机体系结构、分布式系统、云计算等. Email：jiangxh@zju.edu.cn. 吴朝晖，男，1966年生，博士，教授，主要研究领域为服务科学与网格计算、嵌入式普适计算等. Email: wzh@zju.edu.cn. 叶可江，男，1986年生，博士研究生，主要研究方向为虚拟化与云计算、性能评估与建模. Email：yekejiang@zju.edu.cn.

1 引言

随着大规模数据中心在全球范围内的广泛部署，其高能耗、高费用、高污染等问题日益突出[1]。以 Google 公司为例：在过去的十年时间里，其数据中心耗电量增加了 20 倍之多[2]。随着处理器制造工艺的不断进步，Intel 的 Itanium2 处理器集成的晶体管数量达 10 亿个[3]。2010 年数据中心的功耗密度（Power Density）也达到 60Kw/m^2[4]。数据中心的高能耗也导致了许多环境问题。据统计，2007 年全世界的数据中心的二氧化碳排放量几乎和整个阿根廷接近，并保持高达 11%的年增长率[5]。特别是随着云计算的到来，更多的资源集中到云端，给能耗的高效管理带来了更大挑战[1]。如何降低绿色数据中心能耗，构建绿色数据中心（Green Data Center）受到越来越广泛的关注[6]。

传统的节能方式一般是减少节点的计算能耗。例如采用处理器电压频率调整（Dynamic Voltage and Frequency Scaling, DVFS）等底层节能技术；对任务负载进行调度，将任务进行集中化；或采用虚拟化技术，通过服务器整合把多个虚拟机整合到同一个物理机上，关闭空闲的物理机，达到节能目的。这些方法对节能起到了巨大的推动作用，但是都没有考虑到制冷设备的能耗问题。实际上，数据中心的能耗不仅仅来源于服务器的计算能耗（包括处理器、磁盘、网络设备等的能量消耗），还包括制冷能耗。其中制冷开销是最为主要的部分[2][7]，接近所有能耗开销的 50%[8]。我们把将数据中心高制冷能耗的原因总结如下。

1）制冷设备需要将数据中心的温度严格控制在合理范围内。首先，数据中心温度过高会严重影响设备的可靠性。Little Blue Penguin Cluster 的经验数据表明[9]，温度每升高大约 10 度，设备的故障率就会翻倍。为保证设备工作在正常温度下，往往需要消耗大量的制冷能耗。其次，数据中心内部的计算设备能耗高，转化并耗散了大量内能。这给温度控制带来了巨大挑战。以 TACC's Ranger 和 IBM Blue-Gene/L 为例，其冷却率分别为 1:1.5（即每消耗 1.5w 计算功率，需要额外 1w 功率进行冷却）和 1:2.5[4]。

2）数据中心温度分布不均匀。从空间角度来看：由于硬件布局、负载分布等原因，会造成数据中心的温度在不同节点、不同位置分布不均衡；从时间角度分析：数据中心所处的外部环境会因为自然现象周期性地改变，其内部的计算资源利用率会随着用户任务负载的变化而变化。因此，数据中心节点温度的波动要求制冷设备进行过量冷却，即需要确保在最易出现危险的情况下设备也是可靠和安全的。

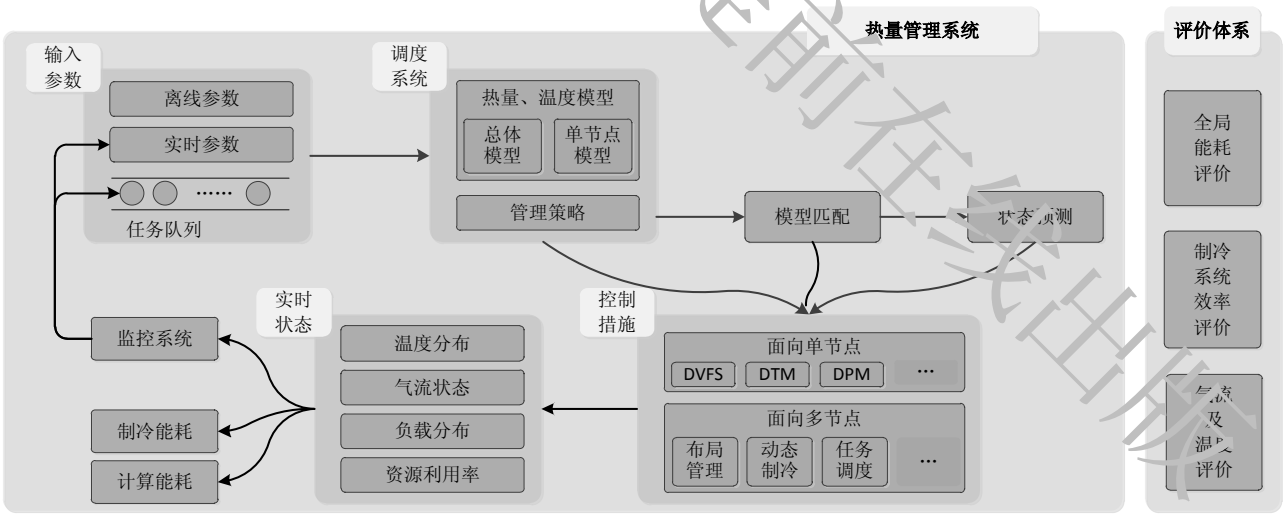

图1 绿色数据中心热量管理架构

3)数据中心的热量管理(Thermal Management)方法存在不足，降低了制冷设备的运行效率。数据中心的热量管理包括对温度、气流等物理参数进行管控的一系列软硬件方法。其主要目的是：1）合理控制温度，提高设备的可靠性；2）优化数据中心的热力学特性，尽量减少热点（Hot Spot）的产生。由于最小化峰值温度与最小化制冷能耗之间具有等价性[10]，因此减少热点有助于提高制冷效率，减少制冷能耗。绿色数据中心的热量管理架构如图 1 所示，主要包括管理系统和评价体系两部分。调

度系统作为管理系统的核心，以用户任务、离线参数和实时参数作为输入，对数据中心进行控制。其中，离线参数包括数据中心的设备布局和热力学参数等；实时参数指数据中心的温度分布、负载分布、系统资源利用率等。如果采用基于预测的管理策略，管理系统需要经过模型匹配和状态预测两个步骤。不同的管理策略和配置会导致不同的数据中心状态。监控系统负责对状态进行实时监控，更新实时参数和用户任务队列，并进入下一个控制循环。评价体系致力于对不同热量管理系统的效果和作用进行评价。评价参数可分为全局能耗评价、制冷系统效率评价、热量及温度评价三部分。

本文从状态监控、热量建模、热量管理策略以及热量管理评价四个方面，对近年来绿色数据中心热量管理领域的研究进展和最新成果进行全面地综述。本文第 2 节介绍了绿色数据中心的监控框架和具体实现。第 3 节分析了数据中心的建模，并重点综述了数据中心的常用数学模型及热力学模型。第 4 节根据面向单节点和面向多节点的分类对现有的热量管理策略进行了总结。在第 5 节中，综述了管理策略的性能指标和评估方法。最后对全文进行总结，并指出值得进一步研究的方向。

2　数据中心监控框架与实现

实时地监控数据中心的各类数据是智能制冷、热量管理、负载迁移等的先决条件[11]。现代数据中心存在大量的网络设备和分布式计算节点，具有耦合性低、可靠性差、经常需要扩展或升级的特点，使得数据中心监控系统的设计和实现面临一定的挑战，需要满足以下特性。1）可伸缩性：监控系统必须能够应对节点数目的动态变化，具有可伸缩性。2）健壮性：即使部分节点产生故障，仍然可以提供有效服务，具有高度的容错性。3）可扩展性：监控的数据类型具有可扩展性，即允许用户添加新的监控数据类型。4）易管理性：对监控系统的管理开销不能随着节点数目的增加而线性增加，尽量减少手动配置的内容。5）可移植性：能够被移植到不同操作系统或芯片架构上。6）系统开销低：为了避免与用户程序发生资源争用，监控程序不能消耗过多系统资源[12]。7）可编程：监控设备符合统一规范，具有标准接口监控程序可以快速、及时、准确地获取设备的测量数据；8）自动化：在无人工干预的情况下自动收集测量数据；9）高可靠性：设备可进行 7×24 小时无人值守的监控，在长时间不间断工作的情况下保证设备的正常测量，并提供必要的预警、报警等应急功能。

本文所讲的监控主要针对两个方面：首先是环境监控，即对数据中心的自然环境和辅助设备进行监控，包括对内部空气的温度、湿度、气流速度等数据的测量与收集；其次是资源监控，即对服务器集群、网络设备等资源的物理参数和使用状态等进行监控。在环境数据的监控方面，主要方法是通过一定的硬件设备来获得各项环境参数。例如，可使用温度计、湿度计、流量计、气流监控器等设备对服务器节点以及数据中心内部气流的各种物理参数进行测量。例如 Shen H 等人提出的基于阵列传感器的数据中心温度场可视化系统[13]，利用分布于数据中心各个位置的传感器所测量的温度作为边界条件。再利用计算流体力学软件进行分析，得到整个数据中心的温度场分布。

与环境监控不同，资源监控指对云计算资源设备的状态参数和运行情况进行监控。我们将其分为三种不同的情况。1）通过计算机设备或操作系统的接口直接获得，例如 CPU 利用率、内存或网络带宽的使用情况等；2）对于部分参数，计算机和操作系统都未提供相应的测量接口，需要通过软硬件结合的方法或间接计算得到。例如 CPU 功耗，文献[14]提出了一种不依赖电学测量仪器，而是通过 CPU 工作频率和利用率进行推算的方法。3）依赖集成在计算机主板、芯片上的传感器进行监控。以温度测量为例，目前大部分的处理器芯片内部、部分计算机主板上都集成了温度传感器。这个传感器将测量的温度数值大多存放在 CPU 寄存器中，通过驱动程序可以进行读取。

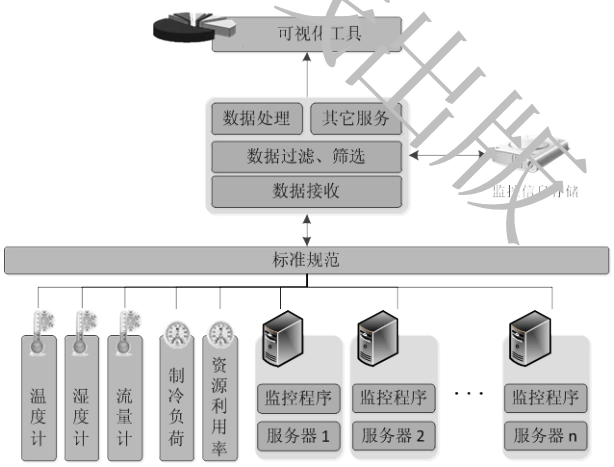

图2　分布式监控系统框架

Moore J 等人开发的自动收集和分析数据的系统[15]可以看成是数据中心监控系统的一个典型。该系统主要包括监控、过滤、分析三个部分：监控部分的设计必须具有高度的可扩展性，以满足数据中心的动态变化以及对测量要求的不断调整；过滤部分使用的基本过滤方法包括去重，以及过滤变化不大的数据；分析部分对数据的行为特征、参数间的相互影响关系进行分析，为调度策略的制定提供参考依据。这种分布式的监控模式是最为常见的一个典型，文献[16][17][18]也使用了类似的监控模式。

通过对诸如Ganglia[12]，Zabbix①，Nagios②等大量监控系统的研究，并结合现代绿色数据中心的特殊性，我们抽象出如图 2 所示的分布式监控系统框架。一般通过特定的测量设备完成环境监控，并在每个服务器上运行监控程序（例如可以是一个守护进程）进行资源监控。这些设备和进程通过一定的标准规范与主节点进行通信，传输收集的数据。主节点对数据进行分析、处理、筛选、存储等操作，必要时还会以某种方式进行报警、通知管理员或采取应急措施等。最后通过可视化工具（可选）对监控数据进行呈现。

3 数据中心热量及温度建模

3.1 数据中心基本温控结构

热量与温度存在直接关系。不同绿色数据中心的规模、结构、布局、制冷设施各不相同，但温度变化的根源都来自于热量的变化。如式(1)所示：

$$\Delta T = \frac{\Delta Q}{cm} \tag{1}$$

其中，ΔT 是温度的变化值，c、m 分别是部件的比热容与质量，ΔQ 是物体的热量变化值。具体到数据中心的部件可分为两类。1）本身会消耗电能的部件：例如计算机、网络、空调等；该类部件的热量变化源自于本身消耗的电能转化而成的热量；另一部分源自于与周围接触物体（如散热片、空气）的热传递。2）其它部件：热量变化完全取决于周围环境的热交换。如果把数据中心看作一个整体，一段时间内平均温度的改变则取决于其产生的热量与制冷设备移除的热量[19]。

$$\Delta T^{room} = \frac{H_{room} - Q_{room}}{M_{room} C_p} = \frac{\Delta I_{room}}{M_{room} C_p} \tag{2}$$

其中，M_{room} 和 C_p 表示数据中心内部总质量和比热容；H_{room} 表示整个数据中心的产热量；Q_{room} 表示制冷系统抽取的热量，两者的差值直接决定室内平均温度的变化。

数据中心有多种供冷回流模式（Supply and Return Scheme）[19][20]。不失一般性，我们以“地板送风，水平回流”的模式为例，描述绿色数据中心的整体温控结构。数据中心下部是地板空层（Raised Floor Plenum），数据中心空调设备（Computer Room Air Conditioning Unit，CRAC）的低温气流通过该空层并经由通风地板（Vent Tile）进行送风。服务器以行为单位放置。地板送风系统的制冷气流从两行机架（Rack）间送出，从机架的前方入口流经机架，带走将服务器耗散的内能之后变成高温气流从机架后方出口排出。一般而言，每个节点内部都安装了各种类型的风扇及冷却组件，如机箱风扇、CPU 风扇等。从机架后部排出气流的密度由于温度升高而降低，受到浮力的作用自然上升并接近水平地返回 CRAC。每两行机架之间，如果直接受到地板送风气流的制冷效果影响，温度便会较其他区域低，形成“冷道”（Cold Aisle），反之在机架的后部，由于热气流的汇聚，温度升高，形成“热道”（Hot Aisle）。

在数据中心的整个制冷系统中，CRAC 扮演着心脏的作用。它不断吸收受热后的气流，并送出低温制冷气流，即不断地将数据中心内部产生的内能“搬运”到外部，同时这个过程也需要消耗一定的能量，也就是制冷能耗。在不同条件下，搬运同样内能所消耗的制冷能耗会有所差异，具体体现为 CRAC 的工作效率，使用 *CoP*（Coefficient of Performance）参数进行表征[8][21]。其定义如下：

$$CoP = Q / W \tag{3}$$

其中，Q 表示移除的能量，W 表示 CRAC 本身消耗的能量。*CoP* 越高代表制冷设备的制冷效率越高，即移除相等的热量所消耗的电能越少。同一个 CRAC 的 *CoP* 值也并不固定，它会随着工作温度变化而变化。一般而言，CRAC 送出的气流温度越低，*CoP* 值也越低，其制冷效率下降。来自惠普实验室数据中心的数据表明，其水冷 CRAC 的 *CoP* 值可以总结为经验公式(4)[10]。其中，T 是 CRAC 的供冷温度设定值。

$$CoP = \left(0.0068T^2 + 0.008T + 0.458\right) \tag{4}$$

同时，数据中心内部设备由于可靠性的要求，

① http://www.zabbix.org
② http://www.nagios.org

需满足如下限制条件[8]：

$$\boldsymbol{T}_{inlet} \le \boldsymbol{T}_{red} \quad (5)$$

$\boldsymbol{T}_{inlet}$ 和 $\boldsymbol{T}_{red}$ 是 n 维向量，表示各机架的入口温度和最高临界温度，n 为机架数量。即对任何机架而言，进入机架的冷却气流温度不能超过一定值。由于数据中心节点温度分布的非均衡性，需要将CRAC的工作温度设定在较低的范围内以保证内部设备的可靠性。*CoP* 值会由于较低的CRAC工作温度设定而消耗更多的制冷能耗。如何在保证设备可靠性的前提下，尽量提高空调设定温度是建立绿色数据中心热量管理体系最主要的目标之一。

3.2 数据中心总体热量及温度建模

数据中心往往结构复杂，造价昂贵，对服务可靠性要求高，因此很多研究和实验没有条件运行在真正的数据中心环境下，而需要采用仿真的办法[20]。此外，很多热量管理策略（详见第4节）都需要对温度进行预估，即预测节点在之后某一时刻的温度。实际上，仿真和预测都依赖于数据中心温度模型的建立。在绿色数据中心的热量管理领域，总体建模是把数据中心整体当成建模对象，需要考虑整体的热力学规律和数学表达。总体建模是面向多节点热量管理策略的基础，主要对数据中心的热学特征、热量/气流循环等问题进行研究。其主要方法是借助机器学习、神经网络、流体力学定律等对数据中心进行不同程度的抽象和概括。

计算流体力学（Computational Fluid Dynamics，CFD）是广泛使用的一种建模方法。它使用数值方法在计算机中对流体力学的控制方程进行求解，从而预测流体的状态[20]。CFD的基本思想是把现实条件下连续的流体划分为离散的格点，用离散的方式使用计算机进行处理。根据需要，可以选择不同的方程对流体进行描述：使用欧拉方程描述粘滞流体，使用纳维-斯托克斯方程描述零粘滞的理想流体。目前，有许多软件可以帮助进行流体建模并求解，如TileFlow①、FloVENT②、Flotherm③、Fluent④等。文献[22]是最早提出使用CFD对数据中心进行建模的工作之一。随后，围绕使用CFD和热传导（Heat Transfer）进行模拟并研究数据中心热力学规律这一课题出现了大量的研究工作。文献[20]对这些工作进行了总结，并将CFD/HT的建模和研究工作分为6种类型：1）对地板送风制冷系统的气流进行模拟并对通风地板的气流速率进行预测；2）研究数据中心硬件设备布局对温度分布和制冷效果的影响；3）探讨不同供冷回流模式的气流和温度特点；4）能耗使用效率及制冷效率的评估；5）单个机架的温度分析；6）考虑服务器负载或CRAC制冷负载动态变化时数据中心的控制和生命周期管理。

使用CFD的方式进行仿真具有较高的准确性，但一般而言，需要花费相当长的时间（小规模的数据中心需要约一个小时才能获得收敛的结果，对于大规模的或者切分格点更多的数据中心而言，所需时间将大大增加[23]）。因此，CFD仅适用于离线情况，而对一些需要即时模拟，快速反应或者在线使用的场景则并不适用。基于此，Tang Q等人将模型进一步抽象化和简单化，牺牲一定的准确性来获取更快的模拟速度[23]。该方法最大的特点是考虑了节点之间的热量交叉影响，并把该影响的强弱抽象为一个 n 阶矩阵 $\boldsymbol{A}$（n 为数据中心的节点数目）。该方法首先通过对不同位置的气流参数及服务器功率进行测量并求解出矩阵 $\boldsymbol{A}$。该矩阵实际上记录了数据中心的热学特征。之后，在获得数据中心实时信息的基础上，根据其构建的模型和计算方法便可预测数据中心的温度分布。但该方法具有以下两点局限性：1）该方法没有考虑时间因素，即预测数据中心在足够长时间之后的一种相对稳定的状态；2）该方法假定冷却气流能够完全并及时地带走服务器耗散在周围空气中的全部热量，而实际并非如此。这两点对都在一定程度上影响预测的准确性。与此类似，Heath T等人在考虑整个数据中心时，用图论的方法将其简化[24]：以顶点代表数据中心的元素实体（如服务器、CRAC等），边代表元素之间的气体流动或热量传输。相对于Tang Q等人的工作，该文献在热量传输和温度变化方面考虑了时间的因素。

数据中心的内部结构与神经网络存在很多相似之处[17]。例如：制冷气流从CRAC流经节点，再回流到通风口，类似于在神经网络中输入值到输出值的计算过程；数据中心节点相互之间存在影响，并且强弱不一，类似于神经网络的节点之间的连接。Moore J等人使用神经网络对数据中心进行建模。对初始化后的神经网络，利用实际监控所获得的数据进行训练，每对数据包括节点功率值（输入）

① http://inres.com/products/tileflow/overview.html
② http://www.mentor.com/products/mechanical/products/flovent
③ http://www.mentor.com/products/mechanical/products/flotherm
④ http://www.ansys.com/Products/Simulation +Technology/Fluid+Dynamics/Fluid+Dynamics+Products/ANSYS+Fluent

和温度分布（输出）。与文献[23]类似，由于不能考虑时间这一关键因素，这种基于神经网络的方法具有一定局限性，只能针对稳定的状态进行预测。

3.3 数据中心组件热量及温度建模

数据中心的组件建模主要研究内部单位（包括服务器、机架、CRAC 设备等）的热学模型，是在总体建模基础上的进一步细化。本文将其分为功耗模型与温度模型两部分进行综述。

3.3.1 组件功耗模型

对于一个典型的服务器，热量来源不仅包括 CPU、还包括 I/O 设备、内存、磁盘和网卡[19]。每一部分的产热量与消耗的功率呈正相关。一般而言，消耗的功率并不会严格等于产热量（例如对于磁盘，部分功率的消耗转化为磁盘转动的机械能）。节点单位时间内的产热量可表示为：

$$h = p^{cpu}\alpha^{cpu} + p^{IO}\alpha^{IO} + p^{mem,stg}\alpha^{mem,stg} + p^{NIC}\alpha^{NIC} \quad (6)$$

α^{cpu}、α^{IO}、$\alpha^{mem,stg}$、α^{NIC} 表示功率到热量的转化系数。其中，CPU 功耗可看作动态功耗、静态功耗和常开单元功耗（Always-on Power Consumption）的总和[25]。

随着云计算的兴起，虚拟化技术被广泛采用以实现不同用户对硬件的共享。对于物理机而言，功率可以通过仪器直接进行测量。而虚拟机的能耗却无法直接进行测量。文献[26]对虚拟机的能耗进行了建模分析，指出虚拟机的能耗可以通过监控物理服务器底层性能计数器的方式来间接获得。一个典型的虚拟机系统的整体功耗可以表示为：

$$P_{total} = P_{baseline} + \sum_{k=1}^{N} P_{domain(k)} \quad (7)$$

其中，P_{total} 表示总体功耗，$P_{baseline}$ 表示空闲时的功耗，$P_{domain(k)}$ 表示第 k 个虚拟机的功耗。$P_{domain(k)}$ 又可表示为各个虚拟机部件的功耗之和，而各个虚拟机部件的功耗则通过硬件性能计数器间接计算得到。此外，Liu 等人对虚拟机动态活动（如在线迁移）的功耗进行了建模[27]。

除了对服务器节点进行热学建模以外，文献[19]给出了制冷设备的功耗模型。CRAC 的能耗可以看成是压缩机和风扇两部分的总和：

$$E_{total} = E_{compressor} + E_{fan} \quad (8)$$

前者与 *CoP* 值直接相关；后者如式(9)所示，与风扇转速的立方成正比。其中，p_{ref} 和 ω_{ref} 是选取的功率和转速的参考值。

$$p = p_{ref}\left(\frac{\omega^3}{\omega_{ref}^3}\right) \quad (9)$$

3.3.2 组件温度模型

对于单个计算节点（如图 3 所示），冷却气流从入口进入（温度为 T_{sup}），流经节点并对其冷却，最后从出口排出并返回 CRAC（温度为 $T_{CRAC,in}$）。图中箭头代表气体的流动。可以看到，从节点排出的热气一部分会再一次重新渗入到本节点和其它节点内部（Heat Recirculation）。类似地，冷却过其它节点的部分热气也会影响该节点，形成交叉影响。另一方面，从供冷处吹出的部分冷气未经过任何节点直接返回 CRAC（Short Circuiting，也叫 By-Passing）[28]。这两种现象在很大程度上影响着 *CoP* 值。很多工作都致力于减少两者的强度，从而提高制冷效率[2][8][10][21]。

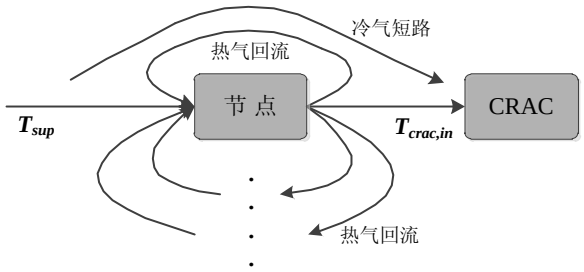

图3 节点交叉热影响模型[29]

一般地，节点入口处的冷却气流温度会受到数据中心内所有 CRAC 的运行状态（包括 T_{sup} 和 CRAC 送风风扇的转速）和热回流强度两者的影响。文献[29]给出了节点入口温度离散化的量化表达：

$$\begin{aligned} \boldsymbol{T}_i(k+1)_i &= \boldsymbol{T}_i(k) + \boldsymbol{F}_i + \boldsymbol{C}_i \\ \boldsymbol{F}_i &= \sum_{j=1}^{N_{CRAC}} \boldsymbol{g}_{i,j}\left[\boldsymbol{T}_{sup,j}(k) - \boldsymbol{T}_i(k)\right] * \boldsymbol{VFD}_j(k) \end{aligned} \quad (10)$$

其中，$\boldsymbol{T}_i(k+1)$和 $\boldsymbol{T}_i(k)$分别是时刻 $k+1$ 和时刻 k 时的入口温度。$\boldsymbol{F}_i$ 和 $\boldsymbol{C}_i$ 分别代表空调和热回流对温度的影响。$\boldsymbol{F}_i$ 是所有空调运行状态参数的加权和。

上述公式给出了节点入口温度的一般形式。对于节点本身的温度，RC 模型是进行精确预测的常用方法。在很多物理场景中，热力学的变量与电学变量存在一一对应的关系。这种对应并不是变量本质上的一致性，而是在模型与计算上存在相似性。例如：热量的传递与电流、温度的差值与电压，热阻与电阻、热容与电容等。数据中心的内部服务器的热学模型可以使用电学中的动态电路模型加以解决。RC 电路（Resistor-Capacitor Circuit）模型就

是把服务器内外看成是具有一定温度差和热阻的热量传输体系。如图 4 所示：T 代表服务器内部温度，T_{amb} 代表服务器外部温度，P 是服务器功率，R 为服务器热阻，C 是服务器比热容。

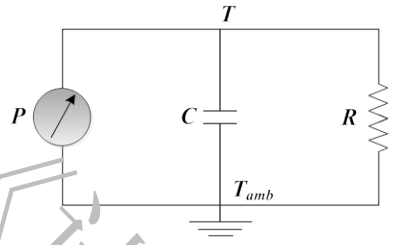

图4　RC 模型[30]

我们考虑一段很短的时间[0，t]，把功率 P 看成是固定值，根据基尔霍夫定律和欧姆定律求得 t 时间后服务器的温度：

$$T = PR + T_{amb} + (T_0 - PR - T_{amb})e^{\frac{-t}{RC}} \quad (11)$$

可以看到，服务器的内部温度随着时间呈指数型变化。如果满足 $T_0 < PR + T_{amb}$，则内部温度将逐渐上升，反之亦然。RC 模型将节点功率、节点内外温度变化联系起来，建立起定量关系。这对温度预测而言，具有非常重要的意义。文献[30][27][31][32]的工作都采用了 RC 传热模型，并且都做了功率 p 与 T_{amb} 在[0，t]内恒定的假设。这些假设虽然简化了微分方程的求解，但都在不同程度上牺牲了预测的准确性。实际上，如果去除该限定，即假设 p 和 T_{amb} 是时间的函数，会增加求解的难度。同时，由于功率以及外部温度随时间变化的函数是难以预计的，使用该假定因此具有相当的困难。

对于数据中心内部的计算机节点，功率是其利用率的函数，即为 $p(Utilization)$。该节点与周围接触物体会进行持续的热传递，从而导致温度的变化。具体通过如下方程组描述：

$$Q_{gained} = Q_{transfer} + Q_{component} \quad (12)$$

$$Q_{transfer,1\to 2} = k \times (T_1 - T_2) \times time \quad (13)$$

$$Q_{component} = p(Utilization) \times time \quad (14)$$

$$p(Utilization) = p_{base} + Utilization \times (p_{\max} - p_{base}) \quad (15)$$

$$\Delta T = \frac{\Delta Q}{mc_0} \quad (16)$$

上述方程组的详细解释见附录 1。与 RC 模型类似，该方程组同样定量的描述了节点与周围环境热传递的规律，并将这些物理量与温度建立了联系。文献[18][24]的工作采用该方法进行温度预测，但是未给出具体的计算过程。实际上，我们将其应用到节点和节点周围环境的场景当中时，经过计算，最终得到和式(11)形式完全统一的结果（见附录 1）。

以上几种温度建模方式都是根据数据中心的热学现象进行简化和抽象，进而对温度进行预测，是先验的理论方式。而文献[33]则从经验总结的角度出发，观察不同的任务在服务器上执行时产生的热量以及引起的温度变化，具体是在单台服务器上执行 SPEC 的基准程序，并总结不同类型任务对服务器温度变化影响的规律。

4　数据中心热量管理策略

绿色数据中心的热量管理策略是在满足一定的约束条件（例如设备温度不能超过一定阈值）下，寻找最优解（任务的调度方案，制冷设备的配置等）的过程。在不同的应用场景下，其目标有所差异。所以对该优化问题的定义也有所不同。根据约束条件的不同，大致可以将热量管理的问题分为两类：1）在所有设备的温度不高过其温度阈值的前提下，减小能耗或增加计算吞吐量；2）在满足用户 QoS(Quality of Service)的前提下，最小化系统能耗。

对于第一类问题：防止设备温度过高是基于硬件的可靠性提出的，在此基础上进一步提出节能和提高计算性能的要求。例如将热量管理问题定义为：通过一定的热量管理方法，保证硬件设备温度不超过预定值，并且最小化制冷能耗[19]或最大化系统计算性能[33][34]。第二类问题则出现在对任务完成时间要求严格的场景下，此时往往把任务的周转时间和截止时间放在首位，在此基础上最小化数据中心能耗[8]。

绿色数据中心的热量管理策略多种多样，根据不同的划分标准可进行不同的分类。例如根据面向的物理机数目分：有针对单节点温度控制的硬件技术和调度方法；有针对数据中心全局的设施布局和调度方法。按照灵活度可分为动态自适应的（这里的动态自适应指的是可以根据当前环境或者历史记录动态调整管理策略或参数，从而具有更好的适应性）和非自适应的等等。图 5 给出了具体的分类方法和相应类别。本节首先引出绿色数据中心热量

管理问题的定义，再对其中具有代表性的管理方法按照面向单节点和面向多节点进行分类介绍，并详细地对各种热量管理策略进行了综合对比（详见表1）。

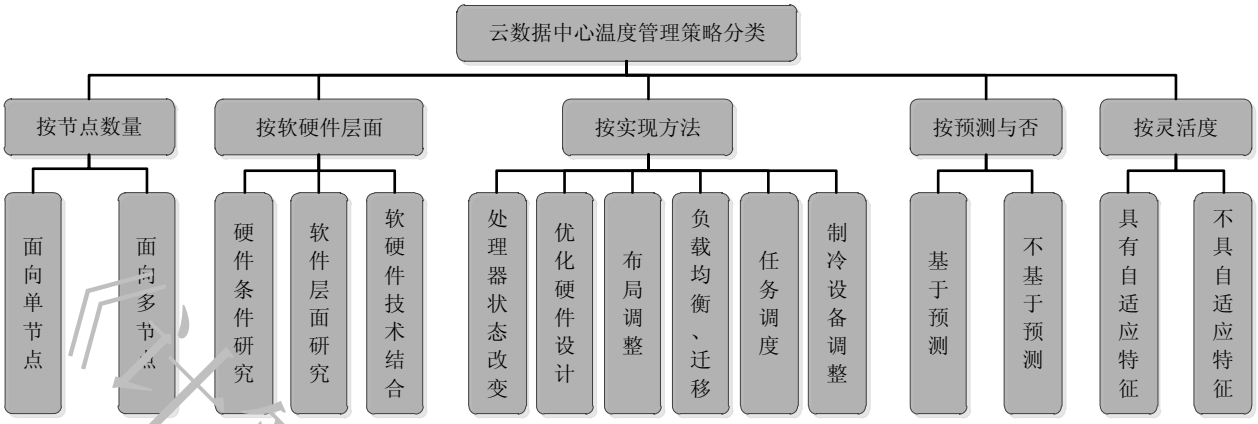

图5 热量管理策略分类

4.1 面向单节点的热量管理

通过 RC 模型可以看到：对于单个节点，设备的功率和温度存在直接联系；另外，单节点不存在温度均衡等优化方法。因此，单节点的热量管理问题在某种程度上可以看成是能耗的管理问题。

传统的面向单个节点的热量管理方法主要是考虑对硬件的优化，或者根据节点的能耗及温度进行动态地调整。例如优化处理器架构进行温度控制，比如：对 CPU 指令寄存器进行限制[35]，使用时钟门控（Clock Gating）技术，控制 CPU 流水线的预测分支[36]等。DVFS[37][38]是常用的 CPU 温度控制技术。通过动态调节 CPU 时钟频率和供电电压，可以有效地调整 CPU 功率，改变其能量消耗。Blem E 等讨论了 CPU 设计时在 RISC 和 CISC 指令集的选择上对能耗的影响[39]。Agrawal A 等提出了 Refrint 方法用以优化 Cache 的刷新机制，从而减少能耗[40]。另外也有大量的针对内存系统的能耗优化工作，例如 Lee Y 等人针对 DRAM 内存提出了 Skinflint 内存系统，通过尽量减少写次数来降低内存能耗[41]等。

动态热量管理（Dynamic Thermal Management, DTM）[36]是基于一系列面向单节点的温度控制技术的总结提出的，是指计算机在运行时、根据部件的实时温度进行动态控制的技术。DTM 的一般运行机制为：当系统的状态达到某个预设条件时（如温度超过阈值），便会触发一定的调节措施；在进行调节的同时，系统会持续监控节点状态，在合适的时候停止调节措施，返回正常运行状态。一个高效的 DTM 系统应满足触发机制简单、调节措施有效、调节策略设计合理等特点。动态功率管理（Dynamic Power Management，DPM）是单节点节能的有效技术。它通过动态地对系统组件进行配置，例如关闭空闲组件、降低组件的运行效率、改变系统运行模式等方法，只提供满足要求的最低运行性能，从而达到节省计算能耗的目的。文献[42]对该领域的工作进行了总结，并将其划分为基于预测的 DPM 方法和基于随机控制（Stochastic Control）的 DPM 方法。

温度反过来会对电子元件功耗产生影响；同时，制冷风扇的工作情况会改变节点的热阻，从而影响节点的散热效率和温度变化[25]。在综合考虑制冷能耗与计算能耗情况下，文献[25]提出了一种动态配置制冷设备及 CPU 工作频率的能耗优化策略。对于多核节点的场景，文献[43]则提出了一种基于预测的热量管理方法 NADTM（Neighbor-Aware Dynamic Thermal Management）。该方法考虑了节点内各个核之间的温度影响，利用核之间的任务迁移和 DVFS 技术，在保证不超过温度阈值的同时，最大化节点计算性能。

CPU 的运行可以划分为几个离散的状态（包括不同的运行状态以及睡眠状态），分别对应不同的功耗。对于一个固定的任务序列，系统应该如何为每个任务选择相应的 CPU 运行状态，才能够在温度不超过阈值的前提下，最小化任务的完成时间。文献[27]指出寻找该问题的最优化调度方式是一个 NP 问题，并提出了一种在多项式时间内的近似算法（Fully Polynomial Time Approximation Scheme，FPTAS）。该优化算法可以在一定的误差范围内，以多项式时间给出优化解答。

相对于硬件层面的 DTM 技术，软件层面的任务调度方法将带来更少的性能损失，达到更好的热量管理效果。文献[44]提出了操作系统级别的动态热量管理方法。文章指出，在一个节点上，对于同样的任务序列，不同的调度顺序会引起不同的温度变化。例如：对于两个任务，更“热”（即引起更高的温度上升）的任务在更“冷”的任务之前调度执行时，最终的温度会更低。基于此，文章提出 ThreshHot 算法，通过修改 Linux 内核，在操作系统级别实现其调度策略，即每次选择最“热”的任务执行。但 ThreshHot 算法针对单个 CPU 核，对于单物理机的多核情况，Qu S 等人提出了针对多核的优化算法 GSA(Greedy Scheduling Algorithm)[45]。

4.2 面向多节点的热量管理

对于单个节点而言，能耗和温度直接相关，对于多节点而言却未必。也就是说，如果采取合适的热量管理策略，更多的计算能耗可能带来更小的峰值温度。因此，合适的热量管理策略对绿色数据中心而言尤为重要。对于具体的管理系统和管理策略，我们将其分为基于设备布局的热量管理策略、基于任务调度的热量管理策略和基于综合控制的热量管理策略。

4.2.1 基于设备布局的热量管理策略

对数据中心设备不同布局进行研究是进行热量管理、提高制冷效率最早使用的方法之一。该领域的工作主要围绕设备的布局、供冷回流模式、制冷动态配置等方面进行。

使用 CFD 对数据中心进行仿真分析的结果显示：即使非常微小的布局改变，也会对数据中心的温度分布产生巨大影响，从而影响制冷能耗[11]。如果以机架入口温度作为温度基准，一般而言，以下因素会对其造成影响[46]。1）通风地板与 CRAC 的距离，离 CRAC 近的通风地板比远的通风地板出口气流速度更小（由于较近的通风地板附近气流速度快，而流体压强会随着速度增加而降低，从而导致通风地板上下侧压强差变小）。2）数据中心底部空层的高度，这会影响内部气体的流动状态，进一步影响通过通风地板的气流速度与机架入口温度。3）数据中心的高度，这会影响机架后方排出热气的回流情况，从而影响机架顶部（由于热气密度较低，会浮于数据中心较高处）的服务器温度。4）不同的冷热道位置，这不仅会影响底部送风的速度，还会影响本身道内空气的静态压强，从而影响热回流的程度。文献[46]对以上因素分布使用 CFD 模拟进行了研究，试图寻找最佳的数据中心设备摆放和硬件布局。文献[47]则着重研究了制冷设备的摆放位置和不同的 CRAC 气流速度对热量管理的影响。

除了数据中心的设备布局以外，制冷效率很大程度上还取决于供冷回流模式的选择[48]。文献[48]对 7 种常用的供冷回流模式进行了比较，从 CFD 的模拟数据来看，表现最佳和最差的两种模式分别是“底部送风、顶部回流”和“顶部送风、底部回流”。前者是由于热气密度较低，顶部回流的方式使其热阻降低，从而使散热更有效；后者则极易发生冷气未经过任何节点直接返回 CRAC 以及冷热气流混合的情况。设备布局和供冷回流模式的研究往往是静态的，即不需要动态、在线、快速地建立模型和反馈信息，因此往往使用 CFD 进行模拟研究。文献[20]则对数据中心的 CFD 建模工作进行了系统化的总结。

研究表明，动态的 CRAC 制冷配置可以有效的降低制冷能耗[2][11][49][50]。一般而言，动态制冷配置系统包括以下三个部分[49]：1）用以收集当前数据中心制冷状态的传感器网络；2）可动态调节参数的制冷设备；3）获得制冷参数对制冷效果的影响。仿真分析[11]及实际真实测量[49][50]的结果都表明：拥有动态制冷调节能力的数据中心，其制冷能耗可以大大降低。文献[49]研究了 CRAC 制冷温度设定、CRAC 风扇转速、底部送风地板的开闭状态对制冷效果的影响。文献[50]则通过动态控制底部送风地板的开合状态，降低制冷能耗约 20%。此外，使用 Air Economizer（根据环境变化，利用数据中心周围空气制冷的设备）、Water Economizer 等动态控制技术也可以有效地节约制冷能耗[51]。

4.2.2 基于任务调度的热量管理策略

温度感知的任务调度是数据中心热量管理的重要研究对象。在绿色数据中心的场景下，任务需要进行实时地调度，调度决策时间十分有限。所以 CFD 等耗时长的建模方式往往不适于此，取而代之的是其它更简单，计算复杂度更低的建模方式。与基于设备布局的热量管理策略不一样，该类管理策略通过任务调度的方式，在软件层面防止热点产生、减少热回流，从而提高制冷效率。

一般而言，基于任务调度的热量管理框架可以概括一个 MAPE 控制环[15]：1）Monitor，监控数据中心的各种环境指数以及运行状态；2）Analyze，分析 Monitor 所搜集的数据，提取行为特征；3）Plan，制订资源调度分配的策略；3）Execute，执行调度。

MAPE 环体现了热量管理的一般生命周期。该领域的大量研究都采用了该框架或其类似框架[52][53]。如图 6 所示，传感器负责收集数据，系统根据所收集的数据进行分析并制订相应的调度策略，最终由调度系统调度数据中心的任务请求。其中，任务负载代表对数据中心的请求任务，可以是 web 请求，高性能计算等应用程序。

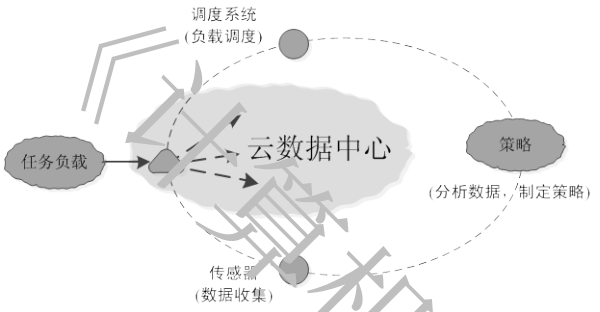

图6 MAPE 控制环[5]

文献[21][54]提出了 6 种基本的温度感知负载调度策略：其思路基本都是均衡各服务器的温度，防止热点的出现，从而降低对 CRAC 的制冷要求，达到节约制冷能耗的目的。

1）均衡负载（Uniform Workload, UW）[21]：调度系统将负载平均的分配到每个节点上。这是最为基本也是最为直观的一种调度方法。

2）最低入口温度调度（Coolest Inlets, CI）[21]：系统对所有服务器进行监控，每次都将负载分配至节点入口温度最低的服务器。该方法需要部署温度测量设备对服务器外围环境进行实时监控。

3）OnePassAnalog[21]：该方法引入了一个功率预算（Power Budget）的概念。它表示调度系统在进行调度时，对某服务器而言，所期望达到的目标功率。调度系统首先根据某种策略计算各服务器的功率预算，然后在负载调度的时候根据该预算进行分配。在具体实现中，调度系统根据下式得出节点的功率预算：

$$\boldsymbol{p}_i = \frac{T_{ref}^{out}}{\boldsymbol{T}_i^{out}} g P_{ref} \tag{17}$$

其中，T_{ref}^{out}、P_{ref} 是选取的参考值（例如可以是整个数据中心的平均值），$\boldsymbol{T}_i^{out}$ 是节点出口的气流温度。可以看到，每个节点的功率预算与节点的出口温度成反比。直观上来说，调度系统希望在高温的节点上分配较低负载，在低温的节点上分配更高的负载，从而均衡节点的温度。

然而，由于 T_{ref}^{out} / $\boldsymbol{T}_i^{out}$ 的取值是连续分布的。因此节点的功率预算值也将是连续的，即可能取一定范围内的任意值。而实际上，要达到并保持一个任意值是非常困难的。若将 CPU 的利用率划分为几个分立的状态：$\{P_1, P_2, \ldots, P_N\}$。其中，$\boldsymbol{P}_i$ 表示节点有 i 个核满负荷工作。可以看到，这些离散的取值难以和 OnePassAnalog 的连续取值相吻合。文献[21]对 OnePassAnalog 方法进行了改进，提出了一种基于区域的离散化方法（Zone-Based Discretization, ZBD）。该方法主要考虑计算密集型的任务，在 OnePassAnalog 的功率预算基础上，进一步以区域为目标进行调整，这种更大粒度的功率调整既保证了与 OnePassAnalog 算法的计算值接近，又增强了节点功率的可控性。

4）均衡出口温度（Uniform Outlet Profile, UOP）[54]。UOP 以均衡所有节点的出口温度为出发点。该方法首先获得节点入口温度的实时测量数据。在假设节点功率与任务的分配具有简单线性关系的基础上，计算出任务的具体调度方法。

5）最小化计算能耗（Minimal Computing Energy, MCE）[54]。该方法尽量减少工作的节点数目，将任务尽量集中地分配在少数几个节点运行。为了减少峰值温度，算法在选择分配时将任务分配到最低入口温度的节点。这种方法着眼于减少计算能耗，而对温度的均衡，制冷能耗的降低没有充分考虑。

6）任务均衡（Uniform Task, UT）[54]，该方法与上述的 Uniform Workload 类似。但是仍然存在区别：Uniform Task 是基于数据中心任务调度实现的；后者则是基于节点功率的调整（即功率预算的概念）。

以上 6 种温度感知的负载调度策略，或者从最小化制冷能耗出发，或者从最小化计算能耗角度出发进行任务调度。这些方法具有直观、简单的优点，但同时过于片面，缺乏对影响数据中心制冷效率的深层原因进行分析。对基于任务调度的热量管理策略而言，增加在制冷效率高的节点上的负载，减少制冷效率低的节点上的负载，是减少热点、提高制冷效率的一个重要思想。因此在进行任务分配之前，需要特定的参数对节点的制冷效率进行评价。例如，可用 *HRF*[21]和 *LWPI*[55]值（详见 5.3 节）表

征某节点的制冷效率。

HRF 是节点产生热回流效果强弱的一个指标。由于热回流是导致冷却效率降低的一个重要原因[21][23][28]，因此 *HRF* 值间接地表征了节点的制冷能力强弱。基于此，文献[21]提出了一种最小化热回流的方法（Minimizing Heat Recirculation, MinHR）。其基本思想是尽量降低 *HRF* 值较低的节点（也就是产生热回流效果强的节点）的功率，增加 *HRF* 值高的节点的功率。MinHR 方法针对功率进行调节而未对任务的具体调度策略进行阐述，文献[4]进行了改进并提出 MinHR-m（Modified MinHR）算法。*LWPI* 则兼顾考虑了节点的热回流效应和空调的制冷能力，用以综合评测空调对某节点的制冷效率。实验表明，基于 *LWPI* 参数所进行任务调度可以在不超过最高临界温度的情况下，节约总能耗的 30%。

HRF 和 *LWPI* 值实际上是为了描述节点的热学特性，根据该特性进行负载调度避免了 CFD 的复杂分析。与此类似，文献[10]提出了基于节点交叉热影响模型（见 3.3.2 节）的 XInt 算法。该模型指出功率消耗的分布情况直接影响到数据中心的气流状态，从而影响温度分布。文献结合任务分配与节点功率的关系，推导出节点温度分布与任务分配的关系并指出不同的分配方案会导致不同的温度分布。XInt 通过遗传算法求出近似最优解（任务调度方案），以保证节点最高温度不会太高（或：最小化最高温度）。文献[4]将该算法重新命名为 XInt-GA 并给出了算法的核心伪代码。除了使用遗传算法以外，该文献还提出了 XInt-SQP 方法，该方法采用序列二次规划（Sequential Quadratic Programming, SQP）解决该非线性规划问题，而 SQP 算法得出的则是实数域内的解答。因此，该方法还需要对得到的解答进行离散化处理。

上述方法（XInt-GA 和 XInt-SQP）所采用的节点交叉热影响模型的最主要贡献是量化了节点之间的相互影响，同时也考虑了节点排出的热气对节点自身的影响，并将这些影响抽象为一个影响因素的矩阵。由于该模型基于以下两个近似假设：1）节点消耗的功率等于产生的热量；2）冷却气流可以及时地带走所有产生的热量。如果节点处于不稳定的状态（例如服务器的功率出现大幅波动），该模型的仿真能力便会下降。另外，该模型重点强调了服务器节点的入口和出口温度，但是忽略了服务器内部温度过高才是导致数据中心可靠性下降的事实。最后需要指出的是，在对任务分配与功率的关系建模中，文献只考虑了单一的任务类型，具有一定的局限性。而文献[33]则进一步对不同类型任务的执行和功耗关系进行了探讨，提出了 Profile-based Temperature-aware Scheduler（PTS）方法。该方法面向大规模集群的任务调度，对每种类型任务的温度上升曲线进行记录，并以此作为调度的参考模型。在具体实现的时候，会对任务的元数据和历史数据进行比对，如果被判断为与历史记录相近，则会使用该历史记录的温度模型作为参考进行调度。如果没有找到相似历史任务，则会重新建立该任务的温度模型，为后续类似任务的调度提供参考。文献[18]着重对提供 web 服务的节点进行热量管理，开发了 C-Oracle。该软件侧重于对不同的应急措施进行分析和预测。这里的“应急措施”指的是服务器温度超过或即将超过阈值等危险状态时所采取的措施，比如 web 请求的重定向、DVFS 等。该系统支持在线预测并对负载进行调度。

对数据中心温度分布的研究表明：数据中心的温度分布是负载分布（*W*）、CRAC 运行状态（*C*）和数据中心布局的函数（*P*）[17]，即：

$$M = T(W, C, P) \tag{18}$$

Weatherman 是一种实时的温度预测方法[17]，使用神经网络的方法对数据中心的功率和温度分布进行研究。该方法具体包括数据采集、模型训练和预测三个步骤。其中，前两个步骤是线下完成的，也就是非实时的。在构建了神经网络的模型之后，再部署到系统中进行实时的预测，帮助调度系统进行任务调度决策。作者进而基于 Weatherman 提出了 Thermal Topology-Based 调度算法。

不同于上述方法中，将整个数据中心看作一个整体并考虑节点间气流的相互影响，Wang L 等人提出的基于实时预测和任务调度的 TASA（Thermal Aware Scheduling Algorithm）、TASA-B（Thermal Aware Scheduling Algorithm with Backfilling）算法[34][52]更为关注单个节点的建模。算法在进行调度时，总是选择最冷的节点执行任务，并尽量先调度“热”的任务，再调度“冷”的任务（见 4.1 节）。为了最大化数据中心的吞吐量，作者在该调度策略的基础上进一步提出了带有回填机制的温度感知调度算法（TASA-B）。某些任务具有多个子任务，需要在多个节点上共同完成，这些子任务一般需要同时开始运行。这可能造成子任务之间的等待，从

而导致资源空闲浪费。所谓回填机制，就是调度系统将任务及时的调度到空闲的节点上，进一步提供数据中心性能。

4.2.3 基于综合控制的热量管理策略

基于综合控制的热量管理策略往往结合布局优化、制冷动态配置、任务调度以及面向单节点的热量管理方法，具有管理全面，节能效果好，实现难度大等特点。基于数据中心综合控制的热量管理方法可分为 Proactive 和 Reactive 两类[19]。前者需要在线下构建一定的模型进行实时评估。后者则基于系统的反馈信息进行控制。例如节点温度高于某一个阈值时，系统便立即进行处理。这种根据反馈信息实时处理的方法相对于 Proactive 方法而言，具有实现简单，反应快速的优势。但是 Reactive 方法也具有阈值难以确定、可靠性低、冷却效率低等缺点。

文献[19]从构建制冷及负载模型的角度出发，提出一种基于动态制冷的热量管理策略。该控制系统通过对数据中心进行热学和流体力学建模，得到制冷设备的不同配置（即工作在不同状态）和制冷能力之间的关系。再根据实时信息对数据中心的产热量进行预估。控制系统在保证设备温度在安全范围内的同时，寻找最佳的制冷配置。HTS（Highest Thermostat Setting）算法综合考虑任务调度与制冷设备的动态控制，是通过综合控制进行热量管理的典型[8]。为了确保节点的温度不超过临界温度（$\boldsymbol{T}_{red}$），需要将 CRAC 的温度设定在一定范围内。而不同的节点由于其物理特征不一样，使得对 CRAC 的温度设定（*CRAC Thermostat Setting*）要求不一样。该方法首先根据 *CRAC Thermostat* 要求对节点排序，并尽量将任务分配到 *CRAC Thermostat* 要求较高（对 CRAC 要求不苛刻）的节点，同时动态调节 CRAC。

表1 热量管理策略的综合比较

策略名称/作者	面向单节点/多节点	软件/硬件	是否基于预测	是否具启发性、自适应性	实现方法	复杂度	实验数据	
							实施效果	数据测量方法
DVFS[37][38]	单节点	软硬件结合	因具体算法而异	因具体算法而异	处理器电压、频率变化	因具体方法而异	节能约 46%	模拟
Refrint[40]	单节点	软硬件结合	×	×	优化刷新策略	较复杂	较使用传统刷新策略的DRAM节能约 47%	SESC
Skinflint[41]	单节点	软硬件结合	×	×	优化内存写操作机制	较复杂	节能约 14%	Zesto[56]
作者：Shin D等[25]	单节点	软硬件结合	√	√	同时控制制冷设备和CPU电压	较复杂	总能耗较baseline的DTM策略减少 8.2%	真实测量
NADTM[43]	多核单节点	软硬件结合	√	×	处理器电压、频率变化	较复杂	不超过临界温度情况下，显著增加处理器吞吐量	真实测量
FPTAS[27]	单节点	软硬件结合	√	×	处理器工作状态调整	较复杂	不超过临界温度情况下，显著增加处理器吞吐量	真实测量
ThreshHot[44]	单节点	软件层面	√	√	任务调度	简单	热量管理效果加强，吞吐量提高 4.1%	真实测量
GSA[13]	单节点	软件层面	√	√	任务调度	简单	热量管理效果显著，吞吐率提高 5.2%~7%	真实测量
作者：Zhou R等[50]	多节点	软硬件结合	√	√	调整底部通风地板开闭状态	较复杂	降低制冷能耗 20%	真实测量
UW[21]	多节点	软硬件结合	×	×	节点功率调节	简单	低利用率时较好	FloVENT
CI[21]	多节点	软件层面	×	×	节点功率调节	简单	高利用率时较好	FloVENT
OnePassAnalog[21]	多节点	软硬件结合	×	×	节点功率调节	简单	各种利用率下表现稳定	FloVENT
ZBD[21]	多节点	软硬件结合	×	×	节点功率调节	较复杂	与OnePassAnalog效果接近	FloVENT
UOP[2][54]	多节点	软件层面	×	×	任务调度	简单	效果一般	FloVENT
MCE[2][54]	多节点	软件层面	×	×	任务调度	简单	在T_{sup}=15℃时，入口温度峰值=30.5℃，较差	FloVENT
UT[2][54]	多节点	软件层面	×	×	任务调度	简单	与UOP接近	FloVENT
MinHR[17][21]	多节点	软硬件结合	×	×	节点功率调整	较复杂	最多可比OnePassAnalog节省制冷能耗 20%	FloVENT
MinHR-m[2]	多节点	软件层面	×	×	任务调度	较复杂	在T_{sup}=15℃时，入口温度峰值=26.3℃，较好	FloVENT

作者：Bash C等[55]		多节点	软硬件结合	×	×	任务调度、节点运行状态调整	较复杂	相对基准方法，节约总能耗约30%	真实测量
XInt	XInt-GA[2][10]	多节点	软件层面	√	×	任务调度	较复杂	两者效果接近，较UT或UOP节省制冷能耗 24-35%	FloVENT
	XInt-SQP[2]								
PTS[33]		多节点	软件层面	√	√	任务调度、迁移	较复杂	有效降低CPU峰值温度	SimGrid
C-Oracle[18]	与Freon集成	多节点	软件	√	√	负载均衡	较复杂	有效应对温度紧急情况	真实测量
	与LiquidN2集成		软硬件			负载均衡、处理器调整			
Thermal Topology Based Approach[17]		多节点	软件层面	√	√	任务调度	较复杂	较UW减少制冷能耗 13-25%	FloVENT
TASA[34][52]		多节点	软件层面	√	√	任务调度	较复杂	较真实数据中心，峰值温度降低 8.9 摄氏℃	根据真实数据中心记录模拟
TASA-B[51]								较真实数据中心，峰值温度降低 8.1 摄氏℃	
Proactive Thermal Management[19]		多节点	软硬件结合	√	√	动态控制制冷设备	较复杂	在保证设备可靠性的基础上，有效降低制冷能耗	MATLAB
HTS[8]		多节点	软硬件结合	√	√	任务调度、CRAC动态调整	较复杂	较baseline算法的SP-EIR值减少15%	FloVENT

5　数据中心热量管理评价

对于不同的绿色数据中心热量管理策略而言，具体的效果和作用需要通过一系列指标进行评定。同时，评价和系统优化有着密切的联系，优化是评价的目的，而评价为优化提供了基础[57]。一般而言，对于数据中心的评价是通过测量设备对系统各项指标直接进行测量或间接计算进行的。绿色数据中心的评价指标必须考虑以下两个因素[58]：1）经济因素，节省数据中心的总运行成本（*Total Cost of Ownership, TCO*[59]）往往是最主要的关注对象；2）环境因素，绿色数据中心的主要目标之一是从节能的角度出发，减少对环境的危害。

本节主要针对绿色数据中心热量管理领域具有代表性的评价指标和评价方法进行综述。我们将绿色数据中心热量管理的评价分为三个方面：1）全局的能耗评价；2）制冷设备的效率评价；3）数据中心的热量及温度评价。我们对这三类评价进行阐述的同时，以表格的形式展现了这些评价的综合比较（见表 2）。

5.1　全局能耗评价

对数据中心的温度进行合理控制可以间接的减少制冷能耗，从而降低数据中心的运行成本。全局的能耗评价是将数据中心看作一个整体，从全局的角度度量数据中心的能量使用效率。这类指标可以总结为数据中心生产率（*Datacenter Productivity, DCP*）[60]，即将数据中心看作一个黑盒，只考虑能耗的输入和有效的产出。该指标表示的是数据中心有效能耗占总能耗的比例：

$$DCP=\frac{Useful\ Work\ Produced}{Total\ Quantity\ of\ a\ Resource\ Consumed} \quad (19)$$

DCP 是一个理论层面的概括，其计算往往十分复杂[60]。具体到能耗的生产率（*Datacenter Energy Productivity, DCeP*），用下式表示：

$$DCeP=\frac{Useful\ Work\ Produced}{Total\ Energy\ Consumed} \quad (20)$$

以数据中心完成任务数为例，在一段时间内（Assessment Window）产出的有用工作（*Useful Work Produced*）为：

$$Useful\ Work\ Produced=\sum_{i=1}^{M}V_i\cdot U_i(t,T)\cdot T_i \quad (21)$$

其中 M 是该时间段内初始化的任务总数、V_i 是一个归一化因子，代表不同任务的权重。如果在该时间段内任务完成，T_i 为 1，否则为 0。$U_i(t, T)$ 是一个基于时间的效应函数，该函数表示随着时间的推移，完成某个任务的价值。文献[60]在数据中心生产率的度量标准上进一步细化，不仅局限于数据中心任务完成数目这个单一指标，而是将其扩展到如网络流量、CPU 利用率、服务器计算能力等多维度的度量中。

与能耗生产率类似，Green Grid①提出了*PUE*（*Power Usage Effectiveness*）和*DCiE*（*Datacenter Infrastructure Efficiency*）用来表征数据中心的功耗使用效率[61]。具体而言，指的是IT设备（包括服务器、存储设备、网络设备、监控节点等）功耗和数据中心总功耗之间的关系。定义为式(22)：

① http://www.thegreengrid.org

$$PUE=\frac{Total\ Facility\ Power}{IT\ Equipment\ Power}$$
$$=\frac{Total\ Facility\ Power\times time}{IT\ Equipment\ Power\times time}=\frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy} \quad (22)$$

DCiE 是 *PUE* 的倒数，即：

$$DCiE=\frac{1}{PUE}=\frac{IT\ Equipment\ Power}{Total\ Facility\ Power}\times 100\% \quad (23)$$

文献[62]进一步对式(22)的分子进行细化，给出了相对具体的计算方法。如果将数据中心 IT 设备的能耗看作有效能耗；则 *PUE*、*DCiE* 反应了数据中心有效能耗占总能耗的比例大小。例如：*PUE* 为 2（*DCiE* 为 50%）的数据中心意味着有效能耗占数据中心总能耗的 50%。据报道，百度南京数据中心的 *PUE* 均值约为 1.3[7]。*PUE* 和 *DCiE* 考虑了 IT 设备的能耗，但没有考虑到有的设备在消耗能量的同时未必会输出有用的工作。例如，两台服务器消耗相近的功率，CPU 的使用率可能相差很大。基于这个原因，文献[60]进一步对其进行了补充，提出一项改进指标 *CPE*（*Compute Power Efficiency*）。

不同的调度算法会产生不同的任务分配和功率变化，由此产生不同的制冷需求。*SP-EIR*（*Energy Inefficiency Ratio of Spatial Scheduling*）指标[8]定义为某算法产生的制冷需求与最佳算法的制冷需求的比值，用以对不同任务调度算法所需的制冷能耗进行评价。

5.2 制冷系统效率评价

数据中心制冷系统效率（*Cooling System Efficiency, CSE*）是制冷系统耗电量与制冷负载的比值，体现了制冷系统运行时的总体效率。

$$CSE=\frac{Average\ Cooling\ System\ Power\ Usage}{Average\ Cooling\ Load} \quad (24)$$

其中，分子是制冷系统消耗的功率；分母是制冷系统的负载（单位是冷吨，1 冷吨表示 1 吨 0℃的水在 24 小时冷冻到 0℃的冰所需要的制冷量）。因此，所谓制冷负载，就是指制冷系统单位时间内运送热量的能力。根据劳伦斯伯克利国家实验室数据中心的实践表明[62]，*CSE* 小于 0.8Kw/ton 是一个较好的标准。

HVAC 系统效率（Heating, Ventilation, and Air Conditioning *System Effectiveness*）[62]与 *PUE*、*DCiE* 类似，是基于数据中心不同组件的能耗比例而言的，它表示 IT 设备能耗与 HVAC 系统能耗的比值：

$$HVAC\ System\ Effectiveness=$$
$$\frac{IT\ Equipment\ Power}{HVAC+\left(Fuel+Steam+Chilled\ Water\right)\times 293} \quad (25)$$

其中，分子和分母分别是 IT 设备的耗电量（Kwh）和 HVAC 系统消耗的能量（包括电能和其它形式能量）。HVAC 系统效率值越大，代表制冷系统的效率越高。当然，由于这是两者的比值，不能单单依据该值的大小断定数据中心的 HVAC 系统的效率或者在不同数据中心间进行比较，而应该进行全面地综合分析。

为了保证数据中心的安全性和可靠性，制冷系统的最高冷却能力不能小于峰值情况下所需的冷却负载。*CSS*（*Cooling System Sizing Factor*）是数据中心冷却系统的总冷却能力与峰值冷却负载的比值。

$$CSS=\frac{Installed\ Chiller\ Capacity}{Peak\ Chiller\ Load} \quad (26)$$

它反映了制冷系统的硬件能力和冷却需求的关系。*CSS* 过高无疑会增加数据中心的 *TCO*；而过低则可能导致冷却能力不足。为了避免温度超过设备的承受范围，*Installed Chiller Capacity* 会适当的大于 *Peak Chiller Load*。根据不同的数据中心布局、热量管理策略、冷却系统的冗余策略等，*CSS* 的值可能变化很大。但理论上，我们应该尽量降低 *CSS* 值

5.3 热量及温度评价

热量和温度状态是绿色数据中心热量管理的最终落脚点。本节主要总结了数据中心常用的热量和温度评测指标，这些指标从不同层次和角度描述了数据中心热量管理的健康程度。

针对单个机架的制冷评价，文献[63]提出了 *RCI*（*Rack Cooling Index*）指标。由于数据中心冷却系统布局以及冷热气体密度不一致等原因，机架不同高度的温度分布（仅考虑入口温度）呈现出不均匀性。若考虑四个温度标准：最大允许温度（*Max Allowable Temperature*, $T_{max\text{-}all}$）、最大推荐温度（*Max Recommended Temperature*, $T_{max\text{-}rec}$）、最小允许温度（*Min Allowable Temperature*, $T_{min\text{-}all}$）、最小推荐温度（*Min Recommended Temperature*, $T_{min\text{-}rec}$）。一般而言，满足如下关系：

$$T_{max\text{-}all}\geq T_{max\text{-}rec}\geq T_{min\text{-}rec}\geq T_{min\text{-}all} \quad (27)$$

如果入口温度高于 $T_{max\text{-}rec}$，代表制冷不足；低

于 $T_{min\text{-}rec}$ 则表示制冷过剩。*RCI* 参数分别考虑了这两种情况，分为 RCI_{HI} 和 RCI_{LO}。

$$RCI_{HI}=\left[1-\frac{\sum_{i=1}^{h}\left(T_i^{in}-T_{\max\text{-}rec}\right)_{T_i^{in}>T_{\max\text{-}rec}}}{\left(T_{\max\text{-}all}-T_{\max\text{-}rec}\right)\times n}\right]\times 100\% \quad (28)$$

将机架从上至下按高度等分地测量 n 个温度样本，T_i^{in} 是第 i 个高度处的机架入口温度。可以看到，如果入口温度的测量位置（以机架高度计算）近似成连续的，则式(28)第二部分的分子是温度超过 $T_{max\text{-}rec}$ 位置处，温度的对高度的积分。对于某个机架而言，分母是一个常数。理想状态下，*RCI* 值为 100%，即没有任何高度位置的温度高于 $T_{\max\text{-}rec}$ 值。与此类似：

$$RCI_{LO}=\left[1-\frac{\sum_{i=1}^{h}\left(T_{\min\text{-}rec}-T_i^{in}\right)_{T_i^{in}<T_{\min\text{-}rec}}}{\left(T_{\min\text{-}rec}-T_{\min\text{-}all}\right)\times n}\right]\times 100\% \quad (29)$$

RCI 反应了单个机架的制冷效果，并通过两个值表征了制冷过剩与制冷不足的总体程度，帮助了解机架的制冷和健康状态。文献[63]在不同制冷模式和功耗密度的情况下对 *RCI* 值进行了研究；文献[64]则对数据中心不同的冷热道封闭模式（包括非封闭，半封闭和全封闭三种模式）对 *RCI* 值的影响进行了探讨。

如 3.3 节所述，影响数据中心制冷效率的两个重要原因是 Heat Recirculation 和 Short Circuiting。两种情况都违反了正常冷却气流的运行方向，严重影响数据中心制冷系统的工作效率。*HRF*（*Heat Recirculation Factor*）参数对单台服务器产生的热回流效应进行了定量地分析。首先，给出热回流效应的计算：

$$\delta Q=\sum_{i=1}^{n}c_p\boldsymbol{m}_i\left(T_i^{in}-T_{sup}\right) \quad (30)$$

其中，n 是节点总数。c_p、$\boldsymbol{m}_i$、T^{in}、$T_{\sup}$ 分别是气体比热、流经节点的冷却气流速度、节点入口气流温度、供冷气流温度（为简化问题，假定所有 CRAC 的供冷气流温度一样）。热回流的根源在于该节点运行时产生了热量。一般而言，产生的热量越多，产生的热回流效应越明显。*HRF* 刻画了节点产生热回流效应的能力强弱。选取数据中心的某个基准状态二元组：$<Q_{ref},\ \delta Q_{ref}>$。其中，前者是数据中心的总产热量，后者是数据中心总热回流效应。然后加大某服务器的功率，并重新测量该二元组的值：$<Q,\ \delta Q>$。*HRF* 定义为：

$$HRF_j=\frac{Q_j-Q_{ref}}{\delta Q_j-\delta Q_{ref}}=\frac{Generated\ Heat}{Generated\ Heat\ \text{Re}circulation} \quad (31)$$

分子是该节点处于两个状态时是产生热量的差值，分母是产生热回流效应的差值。

相较于 *HRF* 值，*LWPI*（*Local Workload Placement Index*）则兼顾综合考虑了节点的热回流效应和空调的制冷能力[55]。该参数定义如下：

$$LWPI_i=\frac{\left(Thermal\ Management\ Margin\right)_i+\left(AC\ Margin\right)_i}{\left(Hot\ Air\ Recirculation\right)_i}=\frac{\left(T_{set}-T_{in}\right)_i+\sum_j\left[\left(T_{SAT}-T_{SAT,min}\right)*TCI_i\right]_j}{\left(T_{in,j}-T_{SAT,i}^{'}\right)} \quad (32)$$

其中，T_{set} 和 T_{in} 分别是节点入口处的温度设定值（例如为最高临界值）和当前温度值；T_{SAT} 和 $T_{SAT,min}$ 分别是空调供给冷气的当前温度和最低温度。TCI_i 是空调 j 对节点 i 的制冷关联系数[55]。分母 $T_{SAT,i}$ 是临近节点 i 的通风地板处的温度。该参数用以综合评价单个节点的制冷效率和制冷“潜力”。

区别于 *RCI*，*HRF* 和 *LWPI* 参数表征单节点的热回流效果，无量纲参数 *Supply Heat Index*（*SHI*）和 *Return Heat Index*（*RHI*）代表了整个数据中心热回流效应的强度[28]。这两个参数从某种程度上可以看成是数据中心热量管理健康状态的指标。用 $\boldsymbol{T}_i^{in}$ 和 $\boldsymbol{T}_i^{out}$ 表示空气经由某节点前后的温度，数据中心共有 n 个节点和 k 个 CRAC。如果整个热循环系统处于稳定状态，则机架散失的热量将等于 CRAC 的制冷量，即：

$$Q_{generated}=Q_{removed}=\sum_{i=1}^{n}c_p\boldsymbol{m}_i\left(T_i^{out}-T_i^{in}\right)=\sum_{i=1}^{k}c_p\boldsymbol{M}_k\left(T_{crac}^{in}-T_{sup}\right) \quad (33)$$

可以看到，通过空调的进出温度 T_{crac}^{in} 和 T_{sup} 可以直接计算整个数据中心的制冷量。两者的差值称作 *Temperature Range*（*RT*）[62]。再根据式(30)，将 *SHI* 和 *RHI* 定义如式(34)、(36)所示。*SHI* 表示了由于热回流导致的能量增加与冷却气体能量增加总

值的关系；*RHI* 则表示 CRAC 吸取的热量与冷却气体能量增加总值的关系。*SHI* 和 *RHI* 都间接的表征了数据中心热回流效应的强弱。如果 *SHI* 越高或者 *RHI* 越低，代表热气回流的现象越严重，反之亦然。

$$SHI = \frac{\delta Q}{Q_{removed} + \delta Q} = \frac{Enthalpy\ Rise\ due\ to\ Infiltration\ in\ Cold\ Aisle}{Total\ Enthalpy\ Rise\ at\ the\ Rack\ Exhaust} \quad (34)$$

$$= \frac{\sum_{i=1}^{n} c_p \boldsymbol{m}_i \left(\boldsymbol{T}_i^{in} - T_{sup}\right)}{\sum_{i=1}^{n} c_p \boldsymbol{m}_i \left(\boldsymbol{T}_i^{out} - T_{sup}\right)} = \frac{\sum_{i=1}^{n} \left(\boldsymbol{T}_i^{in} - T_{sup}\right)}{\sum_{i=1}^{n} \left(\boldsymbol{T}_i^{out} - T_{sup}\right)}$$

$$RHI = \frac{Q_{removed}}{Q_{removed} + \delta Q} = \frac{Total\ Heat\ Extraction\ by\ the\ CRAC\ Units}{Total\ Enthalpy\ Rise\ at\ the\ Rack\ Exhaust} \quad (35)$$

$$= \frac{\sum_{i=1}^{k} c_p \boldsymbol{M}_k \left(T_{crac}^{in} - T_{sup}\right)}{\sum_{i=1}^{n} c_p \boldsymbol{m}_i \left(\boldsymbol{T}_i^{out} - T_{sup}\right)} = \frac{\sum_{i=1}^{n} \left(\boldsymbol{T}_i^{out} - \boldsymbol{T}_i^{in}\right)}{\sum_{i=1}^{n} \left(\boldsymbol{T}_i^{out} - T_{sup}\right)}$$

$$RHI = 1 - SHI \quad (36)$$

文献[65]提出了 *RTI*（*Return Temperature Index*）参数，该参数将气流的 Heat Recirculation 和 Short Circuiting 的两个现象进行了综合。具体为：

$$RTI = \frac{T_{crac}^{in} - T_{sup}}{\left(\boldsymbol{T}_i^{out}\right)_{mean} - \left(\boldsymbol{T}_i^{in}\right)_{mean}} \quad (37)$$

$\left(\boldsymbol{T}_i^{out}\right)_{mean}$ 和 $\left(\boldsymbol{T}_i^{in}\right)_{mean}$ 分别表示节点出入口温度的平均值。理想情况下，应该满足如下关系：

$$T_{crac}^{in} = \left(\boldsymbol{T}_i^{out}\right)_{mean}, T_{\sup} = \left(\boldsymbol{T}_i^{in}\right)_{mean} \quad (38)$$

也就是说，理想状态下，*RTI* 的值是 100%。如果 *RTI* 的值大于 100%，代表热回流效应过强，反之 Short Circuiting 效应过强。文献[65]对不同数据中心结构和不同功耗密度情况下 *RTI* 值进行了研究。结果表明将数据中心的“冷道”和“热道”进行隔离，是减少热回流的有效方法。

为了避免出现单个节点或少数节点过热的现象，往往需要对节点的温度进行均衡。数据中心不同位置的温度方差（*Coefficient of Variance, CoV*）是衡量温度均衡程度的常用指标[54][21]。如果数据中心不同节点的温度相差很大，会导致温度在空间上有较大范围的波动。该参数也会随之增大。

表2 热量管理评价总结与比较

分类	名称		所表征的性质	计算方法	复杂度
全局能耗评价	*DCP*[60]		总体资源利用效率	$\frac{Useful\ Work\ Produced}{Total\ Quantity\ of\ a\ Resource\ Consumed}$	简单
	DCeP[60]		总体能耗利用效率	$\frac{Useful\ Work\ Produced}{Total\ Energy\ Consumed}$	简单
	PUE/DCiE[61]		IT设备功耗比例	$PUE = \frac{Total\ Facility\ Power}{IT\ Equipment\ Power}, DCiE = \frac{1}{PUE}$	简单
	CPE[60]		IT设备利用率与IT设备功耗比例的关系	$\frac{(IT\ Equipment\ Utilization \times IT\ Equipment\ Power)}{Total\ Facility\ Power}$	较复杂
	SP-EIR[8]		算法对制冷效率的影响	$\frac{Energy\ Consumption\ of\ an\ Algorithm\ Under\ All\ CRAC\ Modes}{Energy\ Consumption\ of\ the\ Optimal\ Algorithm\ Under\ All\ CRAC\ Modes}$	简单
制冷系统效率评价	*CoP*[8][21]		制冷设备工作效率	$\frac{Heat\ Removed\ by\ CRAC}{Energy\ Consumed\ by\ CRAC}$	简单
	CSE[58][62]		制冷设备工作效率	$\frac{Average\ Cooling\ System\ Power\ Usage}{Average\ Cooling\ Load}$	简单
	HVAC System Effectiveness[58][62]		HVAC系统能耗与IT设备能耗关系	$\frac{Average\ Cooling\ System\ Power\ Usage}{Average\ Cooling\ Load}$	较复杂
	CSS[58][62]		制冷冗余程度	$\frac{Installed\ Chiller\ Capacity}{Peak\ Chiller\ Load}$	简单
热量及温度评价	*RCI*[63]	RCI_{HI}	机架制冷不足的程度	$RCI_{HI} = \left[1 - \frac{\sum_{i=1}^{h}\left(\boldsymbol{T}_i^{in} - T_{\text{max-rec}}\right)_{\boldsymbol{T}_i^{in} > T_{\text{max-rec}}}}{\left(T_{\text{max-all}} - T_{\text{max-rec}}\right) \times n}\right] \times 100\%$	较复杂
		RCI_{LO}	机架制冷过剩的程度	$RCI_{LO} = \left[1 - \frac{\sum_{i=1}^{h}\left(T_{\text{min-rec}} - \boldsymbol{T}_i^{in}\right)_{\boldsymbol{T}_i^{in} < T_{\text{min-rec}}}}{\left(T_{\text{min-rec}} - T_{\text{min-all}}\right) \times n}\right] \times 100\%$	较复杂
	HRF[21]		单节点产生热回流效应的强弱	$\frac{Generated\ Heat}{Generated\ Heat\ Recirculation}$	较复杂

	$LWPI$[55]	单节点的制冷效率和制冷“潜力”	$\frac{(Thermal\ Management\ Margin)+(AC\ Margin)}{(Hot\ Air\ Recirculation)}$	较复杂
	RT[62]	数据中心总体制冷量	$Return\ Temperature\ of\ CRAC - Supply\ Temperature\ of\ CRAC$	简单
	SHI[28]	整个数据中心热回流效应强弱	$\frac{Total\ Heat\ Extraction\ by\ the\ CRAC\ Units}{Total\ Enthalpy\ Rise\ at\ the\ Rack\ Exhaust}$	较复杂
	RHI[28]	整个数据中心热回流效应强弱	$\frac{Enthalpy\ Rise\ due\ to\ Infiltration\ in\ Cold\ Aisle}{Total\ Enthalpy\ Rise\ at\ the\ Rack\ Exhaust}$	较复杂
	RTI[65]	热回流效应和Short Circuiting效应的综合强弱程度	$\frac{Return\ Air\ Temperature - Supply\ Air\ Temperature}{Rack\ Outlet\ Mean\ Temperature - Rack\ Inlet\ Mean\ Temperature}$	较复杂
	CoV[21][54]	温度分布均匀程度	$Variance\ of\ Temperatures$	简单

6　总结与展望

本文从绿色数据中心的环境与资源监控、热量及温度建模、热量管理策略、热量管理评价四个方面对数据中心的热量管理研究工作进行了综述。文章着重对数据中心的热量管理策略进行了分析总结，从面向单节点/多节点、实现方法、复杂度、灵活度等多个维度对现有热量管理策略进行了归类、比较和分析，并阐述了各种方法的优势和局限性，提出了面向多节点的热量管理机制的理论框架，为绿色数据中心热量管理的后续研究奠定了基础。最后，本文将绿色数据中心的热量管理评价及指标分为全局能耗评价、制冷系统效率评价、热量及温度评价三类进行归类分析和总结。

数据中心的热量管理虽然可以追溯到本世纪初，但是全面的、多节点的综合热量管理策略和方法并不成熟。特别是由于云计算的兴起，基于云计算平台的设备规模大、可扩展性高、功耗密度大、可靠性要求高等特点，数据中心的热量管理面临着新的挑战。我们通过对已有工作的总结，并结合自己的理解，给出未来绿色数据中心的热量管理需要进一步研究的问题。

1）建立统一的物理机温度监控规范。在数据中心的环境下，存在大量异构的物理节点。其内部的温度检测主要依赖于集成在主板和芯片内部的传感器，这些传感器的位置、精度、对应标准的会极大的影响监控系统读数的获取。而该问题的解决需要依靠统一监控标准和规范的建立。

2）考虑监控软件本身的开销和对温度的影响。绿色数据中心的热量管理需要收集监控数据，其中大部分依赖于节点上配置的硬件和软件。软件的运行本身就占用一定开销，同时也会对温度造成影响，而很多前人的工作都忽略了这一点。该问题同时也对轻量级的在线监控技术提出了要求。

3）热量管理与计算能耗的综合考虑与权衡。现有的热量管理方法往往通过负载均衡实现数据中心温度的均匀分布以降低制冷能耗；而从减少计算能耗的角度来看，又往往需要将负载集中化。因此有必要研究如何综合权衡计算能耗与制冷能耗，从而达到最佳的节能效果。

4）针对虚拟化主机的温度建模。随着云计算的兴起，通过虚拟机迁移实现服务器整合以及节能管理将成为未来的主流，因此对虚拟机的温度建模显得尤为重要。以往的温度建模大多停留在对任务、负载及服务器功率的分配和调节上，而缺少面向虚拟机的温度建模。由于虚拟机运行在物理机上时，对于资源的利用率（如 CPU 利用率）并没有一个简单的线性关系。因此，以虚拟机为粒度进行温度建模，并综合考虑虚拟机上不同工作负载的特征，如 CPU 密集、存储密集、I/O 密集等等，是非常必要的，对绿色数据中心节能管理具有重要意义。

5）面向虚拟机的热量管理机制研究。虚拟化技术因具有高可用性、灵活部署、低管理开销等诸多方面的优点，成为绿色数据中心的重要技术。数据中心的任务调度和迁移大多采用虚拟化技术进行，因此，非常有必要对面向虚拟机的热量管理机制进行研究。

6）建立针对云环境下任务特征的温度模型。不同的任务类型在运行时对温度的影响各不相同。例如网站服务、搜索引擎、数据检索的任务往往时间较短，频率较高；而高性能计算往往时间较长、资源利用率高。在此基础上，结合云计算环境下的虚拟化场景，建立任务的温度模型，是很有意思的问题。

7）考虑异构环境的多节点热量管理策略研究。大多数面向多节点的热量管理策略为将问题简化，只考虑同构的数据中心，即各物理节点结构一致。随着高扩展性、动态、异构成为主要绿色数据中心

的新趋势，这些方法很大程度还有继续研究和提升的空间。

8）多维度的热量管理策略研究。目前的工作一般只考虑服务器的能耗和制冷能耗两个部分。单就服务器而言，在大多数情况下，也仅仅是考虑到 CPU 的能耗。在接下来的热量管理研究发展中，强调更加全面地综合管理，会取得更好的效果。

9）将云计算环境下的应用场景与传统的 CFD 建模技术相结合。传统的 CFD 建模主要考虑静态的情况，即数据中心的布局、功率保持相对稳定。这已不再直接适用于绿色数据中心现状。在 CFD 建模时考虑绿色数据中心的多租户、多任务、动态性是非常有意义的研究。

10）建立适用性强、多维度的评价体系。目前数据中心的能耗及温度评价方法和机制大多相对片面或仅考虑数据中心的某个单独方面。有必要建立整套的体系化的评价方法，用于不同数据中心的综合评定和比较。

参 考 文 献

[1] Deng Wei, Liu Fang-Ming, Jin Hai, Li Dan, Leveraging renewable energy in cloud computing datacenters: state of the art and future research, Chinese Journal of Computers, 2013, 36(3): 582-598 (in Chinese)
(邓维, 刘方明, 金海, 李丹. 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, 36(3): 582-598)

[2] Li C, Zhou R, Li T. Enabling distributed generation powered sustainable high-performance data center//Proceedings of the 19th IEEE International Symposium on High Performance Computer Architecture, Shenzhen, China, 2013: 35-46

[3] Lee Y C, Zomaya A Y. Energy efficient utilization of resources in cloud computing system. The Journal of Supercomputing, 2012, 60(2): 268-280

[4] Tang Q, Gupta S, Varsamopoulos G. Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: a cyber-physical approach. IEEE Transactions on Parallel and Distributed Systems, 2008, 19(11): 1458-1472

[5] Kaplan J M, Forrest W, Kindler N. Revolutionizing data center energy efficiency. New York, USA: McKinsey & Company, technical report: 2008

[6] Liu L, Liu X, Jin X, He W, Wang Q, Chen Y. GreenCloud: a new architecture for green data center//Proceedings of the 6th International Conference Industry Session on Autonomic Computing and Communications Industry Session, New York, USA, 2009: 29-38

[7] Li K. Energy efficient scheduling of parallel tasks on multiprocessor computers. The Journal of Supercomputing, 2012, 60(2): 223-247

[8] Banerjee A, Mukherjee T, Varsamopoulos G, et al. Cooling-aware and thermal-aware workload placement for green HPC data centers//Proceedings of Green Computing Conference, Chicago, USA, 2010: 245-256

[9] Hsu C, Feng W, Archuleta J. Towards efficient supercomputing: a quest for the right metric//Proceedings of the IEEE International Parallel and Distributed Processing Symposium , Denver, USA, 2005

[10] Tang Q, Gupta S, Varsamopoulos G. Thermal-aware task scheduling for data centers through minimizing heat recirculation//Proceedings of IEEE International Conference on Cluster Computing, Austin, USA, 2007: 129-138

[11] Patel C, R. Sharma R, Bash C, Beitelmal A. Thermal considerations in cooling large scale high compute density data centers//Proceedings of the 8th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, San Diego, USA, 2002: 767-776

[12] Sacerdoti F, Katz M, Massie M, Culler D, Wide area cluster monitoring with ganglia//Proceedings of the IEEE International Conference on Cluster Computing, Hong Kong, China, 2003: 289-298

[13] Shen Hua-Feng, The design of temperature field virtualization system for internet data center based on the array sensor [MS thesis]. Hangzhou, Zhejiang University, 2013 (in Chinese)
(沈华峰. 基于阵列传感器的数据中心温度场可视化系统研究 [硕士学位论文]. 杭州, 浙江大学, 2013)

[14] Song Jie, Li Tian-Tian, Yan Zhen-Xing, Na Jun, Zhu Zhi-Liang, Energy-Efficiency Model and Measuring Approach for Cloud Computing, Journal of Software, 2012, 23(2): 200-214 (in Chinese)
(宋杰, 李甜甜, 闫振兴, 那俊, 朱志良. 一种云计算环境下的能效模型和度量方法. 软件学报, 2012, 23(2): 200-214)

[15] Moore J, Chase J, Farkas K, Ranganathan P. Data center workload monitoring, analysis, and emulation//Proceedings of the 8th Workshop on Computer Architecture Evaluation Using Commercial Workloads, San Francisco, USA, 2005

[16] Heath T, Centeno A, George P, Ramos L, Jaluria Y, Bianchini R. Mercury and Freon: temperature emulation and management for server systems//Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems, New York, USA, 2006: 106-116

[17] Moore J, Chase J S, Ranganathan P. Weatherman: Automated, online and predictive thermal mapping and management for data centers//Proceedings of the IEEE International Conference on Autonomic Computing, Dublin, Ireland, 2006: 155-164

[18] Ramos L, Bianchini R. C-Oracle: Predictive thermal management for data centers//Proceedings of the 14th IEEE International Symposium on High Performance Computer Architecture, Salt Lake City, USA, 2008: 111-122

[19] Lee Y, Kulkarni I, Pompili D, Parashar M. Proactive thermal management in green datacenters. The Journal of Supercomputing, 2012, 60(2): 165-195

[20] Rambo J, Joshi Y. Modeling of data center airflow and heat transfer:

State of the art and future trends. Distributed and Parallel Databases, 2007, 21(2-3): 193-225

[21] Moore J, Chase J, Ranganathan P, Sharma R. Making scheduling "cool": temperature-aware workload placement in data centers//Proceedings of the 2005 USENIX Annual Technical Conference, Anaheim, USA, 2005

[22] Kang S, Schmidt R R, Kelkar K M, et al. A methodology for the design of perforated tiles in raised floor data centers using computational flow analysis. IEEE Transactions on Components and Packaging Technologies, 2001, 24(2): 177-183

[23] Tang Q, Mukherjee T, Gupta S K S, et al. Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters//Proceedings of the 4th International Conference on Intelligent Sensing and Information, Bangalore, India, 2006: 203-208

[24] Heath T, Centeno A P, George P, et al. Mercury and Freon: temperature emulation and management for server systems//Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems. New York, USA, 2006: 106-116

[25] Shin D, Chuang S, Chuang E, Chang N. Energy optimal dynamic thermal management computation and cooling power co-optimization, IEEE Translations on Industrial Informatics, 2010, 6(3): 340-351

[26] Ye Ke-Jiang, Wu Zhao-Hui, Jiang Xiao-Hong, He Qin-Min, Power management of virtualized cloud computing platform, Chines Journal of Computers, 2012, 35(6): 1262-1285 (in Chinese)
(叶可江, 吴朝晖, 姜晓红, 何钦铭. 虚拟化绿色数据中心的能耗管理. 计算机学报, 2012, 35(6): 1262-1285)

[27] Liu H, Xu C Z, Jin H, et al. Performance and energy modeling for live migration of virtual machines//Proceedings of the 20th International Symposium on High Performance Distributed Computing, San Jose, USA, 2011: 171-182

[28] Sharma R K, Bash C E, Patel C D. Dimensionless parameters for evaluation of thermal design and performance of large-scale data centers//Proceedings of the 8th ASME/AIAA Joint Thermophysics and Heat Transfer Conference, Saint Louis, USA, 2002: p.3091

[29] Zhou R, Wang Z, Bash C E, et al. Data center cooling management and analysis-a model based approach//Proceedings of 28th Annual IEEE Semiconductor Thermal Measurement and Management Symposium, San Jose, USA, 2012: 98-103

[30] Zhang S, Chatha K S. Approximation algorithm for the temperature-aware scheduling problem//Proceedings of the IEEE/ACM ACM International Conference on Computer-Aided Design, San Jose, USA, 2007: 281-288

[31] Skadron K, Abdelzaher T, Stan M R. Control-theoretic techniques and thermal-RC modeling for accurate and localized dynamic thermal management//Proceedings of the 8th IEEE International Symposium on High Performance Computer Architecture, Boston, USA, 2002: 17-28

[32] Rosinger P, Al-Hashimi B, Chakrabarty K. Rapid generation of thermal-safe test schedules//Proceedings of the Conference on Design, Automation and Test in Europe, Munich, Germany, 2005: 840-845

[33] Vanderster D C, Baniasadi A, Dimopoulos N J. Exploiting task temperature profiling in temperature-aware task scheduling for computational clusters//Advances in Computer Systems Architecture, Springer Berlin, 2007: 175-185

[34] Wang L, Von Laszewski G, Dayal J, et al. Towards thermal aware workload scheduling in a data center//Proceedings of the IEEE 10th International Symposium on Pervasive Systems, Algorithms, and Networks, Taiwan, China, 2009: 116-122

[35] Sanchez H, Kuttanna B, Olson T, et al. Thermal management system for high performance PowerPC microprocessors//Proceedings of the IEEE Compcon, San Jose, USA, 1997: 325-330

[36] Brooks D, Martonosi M. Dynamic thermal management for high-performance microprocessors//Proceedings of the 7th IEEE International Symposium on High Performance Computer Architecture, Monterrey, Mexico, 2001: 171-182

[37] Burd T D, Pering T A, Stratakos A J, et al. A dynamic voltage scaled microprocessor system. IEEE Journal of Solid-State Circuits, 2000, 35(11): 1571-1580

[38] Giorgio L, Walter L, Samee U, et al. An overview of energy efficiency techniques in cluster computing systems. Cluster Computing, 2012, 16(1): 1-13

[39] Blem E, Menon J, Sankaralingam K. Power struggles: revisiting the RISC vs. CISC debate on contemporary ARM and x86 architectures//Proceedings of the 19th IEEE International Symposium on High Performance Computer Architecture, Shenzhen, China, 2013

[40] Agawal A, Jain P, AnsariA, Torrellas J. Refrint: intelligent refresh to minimize power in on-chip multiprocessor cache hierarchies//Proceedings of the 19th IEEE International Symposium on High Performance Computer Architecture, Shenzhen, China, 2013

[41] Lee Y, Kim S, Hong S, Lee J. Skinflint DRAM system minimizing DRAM chip writes for low power//Proceedings of the 19th IEEE International Symposium on High Performance Computer Architecture, Shenzhen, China, 2013

[42] Benini L, Bogliolo A, De Micheli G. A survey of design techniques for system-level dynamic power management. IEEE Transactions on Very Large Scale Integration Systems, 2000, 8(3): 299-316

[43] Liu G, Fan M, Quan G, et al. On-Line predictive thermal management under peak temperature constraints for practical multi-core platforms. Journal of Low Power Electronics, 2012, 8(5): 565-573

[44] Yang J, Zhou X, Chrobak M, et al. Dynamic thermal management through task scheduling//Proceedings of IEEE International Symposium on Performance Analysis of Systems and software, Austin, USA, 2008: 191-201

[45] Qu S, Zhang M, Liu G, Liu T. Dynamic thermal management by greedy scheduling algorithm. Journal of Central South University, 2012, 19: 193-199

[46] Bhopte S, Agonafer D, Schmidt R, et al. Optimization of data center room layout to minimize rack inlet air temperature. Journal of

Electronic Packaging, 2006, 128(4): 380-387

[47] Bedekar V, Karajgikar S, Agonafer D, et al. Effect of CRAC location on fixed rack layout//Proceedings of the 10th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronics Systems, San Diego, USA, 2006: 421-425

[48] Shrivastava S, Sammakia B, Schmidt R, et al. Comparative analysis of different data center airflow management configurations//proceedings of the ASME InterPack Conference 2005, San Francisco, USA, 2005

[49] Boucher T, Auslander D, Bash C, et al. Viability of dynamic cooling control in a data center environment//Proceedings of the 9th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, Las Vegas, USA, 2004: 593-600

[50] Zhou R, Bash C, Wang Z, et al. Data center cooling efficiency improvement through localized and optimized cooling resources delivery//Proceedings of ASME 2012 International Mechanical Engineering Congress & Exposition, Houston, USA, 2012

[51] Rumsey Engineers. High performance data centers: a design guidelines sourcebook. San Francisco: Pacific gas and electric company, technical report, 2006

[52] Wang L, Khan S U, Dayal J. Thermal aware workload placement with task-temperature profiles in a data center. The Journal of Supercomputing, 2012, 61(3): 780-803

[53] Mukherjee T, Tang Q, Ziesman C, et al. Software architecture for dynamic thermal management in datacenters//Proceedings of IEEE 2nd International Conference on Communication Systems Software and Middleware, Bangalore, India, 2007: 1-11

[54] Tang Q, Gupta S, Stanzione D, Cayton P. Thermal-aware task scheduling to minimize energy usage of blade server based datacenters//Proceedings of the 2nd IEEE International Symposium on Dependable, Autonomic and Secure Computing, Indianapolis, USA, 2006: 195-202

[55] Bash C, Forman G. Cool Job Allocation: Measuring the Power Savings of Placing Jobs at Cooling-Efficient Locations in the Data Center//Proceedings of USENIX Annual Technical Conference, Santa Clara, USA, 2007: 140-151

[56] Loh G H, Subramaniam S, Xie Y. Zesto: A cycle-level simulator for highly detailed microarchitecture exploration//Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software, Boston, USA, 2009: 53-64

[57] Lin C, Tian Y, Yao Min. Green network and green evaluation: mechanism, modeling, and evaluation. Chinese Journal of Computers, 2011, 24(4): 593-612 (in Chinese)
(林闯，田源，姚敏。绿色网络和绿色评价：节能机制、模型和评价。计算机学报，2011,24(4)：593-612)

[58] Wang L, Khan S U. Review of performance metrics for green data centers: a taxonomy study. The Journal of Supercomputing, 2011: 1-18

[59] Patterson MK, Costello DG, Grimm PF, Loeffler M. Data center TCO: a comparison of high density and low-density spaces. Santa Clara, USA: Intel Corp, Technical report, 2007

[60] Anderson D, Cader T, Darby T, et al. A framework for data center energy productivity. The Green Grid White Paper, 2008

[61] Rawson A, Pfleuger J, Cader T. Green grid data center power efficiency metrics: PUE and DCIE. The Green Grid White Paper, 2008

[62] Mathew P. Self-benchmarking Guide for Data Centers: Metrics, Benchmarks, Actions. Lawrence Berkeley National Laboratory, Berkeley, USA, 2010

[63] Herrlin M K. Rack cooling effectiveness in data centers and telecom central offices: The rack cooling index (RCI). Transactions American Society of Heating Refrigerating and Air Conditioning Engineers, 2005, 111(2): 725

[64] Herrlin M K. Airflow and cooling performance of data centers: two performance metrics. Transactions American Society of Heating Refrigerating and Air Conditioning Engineers, 2008, 114(Part 2): 182-187

[65] Herrlin M K. Improved data center energy efficiency and thermal performance by advanced airflow analysis//Proceedings of Digital Power Forum. San Francisco, USA, 2007

附录 1.

方程(12)-(16)中，等式(12)体现了能量守恒定律：物体的能量变化来自于两个方面，一是与外界的能量交换，二是自身产生的热量。前者的定量描述可以使用牛顿冷却定律计算，如式(13)所示，两个物体之间的能量传递值与两者的温度差和时间成正比。以数据中心服务器节点为例：$Q_{transfer}$ 表示服务器与周围空气之间的热传导值，它与服务器内外温差（T_1-T_2）和时间（*time*）成正比；如式(14)所示，$Q_{component}$ 可以用功率与时间的乘积来计算。功率则使用式(15)，根据服务器利用率进行线性估计。最终由等式(16)给出温度变化的计算公式。其中，c_0 是物体比热。

该方程组描述了两物体之间的热量传递关系，将其应用到服务器（物体 1）和其周围空气（物体 2）中，对服务器温度随时间的改变进行求解，即：

$$T_{node} = T(t) \tag{39}$$

考虑在短时间t内，服务器利用率和外部温度恒定，分别为T_{amb}和p_0，服务器初始温度为T_0。则方程(12)-(16)变为如下方程组：

$$
\begin{aligned}
&Q_{gained} = Q_{transfer} + Q_{component} \\
&Q_{transfer,1\to 2} = \int_0^t k \times \left(T_{node} - T_{amb}\right) dt \\
&Q_{component} = \int_0^t p_0 dt = p_0 t \\
&\Delta T = \frac{Q_{gained}}{mc_0} = T_{node} - T_0
\end{aligned} \tag{40}
$$

整理得到：

$$
\begin{aligned}
&c_0 m\left(T_{node} - T_0\right) = -\int_0^t k \times \left(T_{node} - T_{amb}\right) dt + p_0 t \\
\Leftrightarrow\ &c_0 m T_{node} - c_0 m T_0 - \left(p_0 + kT_{amb}\right) t = -k\int_0^t T_{node} dt
\end{aligned} \tag{41}
$$

求导得到一阶线性微分方程：

$$
\frac{dT_{node}}{dt} = -\frac{k}{c_0 m} T_{node} + \frac{p_0 + kT_{amb}}{c_0 m} \tag{42}
$$

求解得：

$$
T_{node} = C' e^{-\frac{kt}{cm}} + \frac{p_0}{k} + T_{amb} \tag{43}
$$

带入式(41)中，求得：

$$
C' = T_0 - \frac{p_0}{k} - T_{amb} \tag{44}
$$

最终得到：

$$
T_{node} = T(t) = \left(T_0 - \frac{p_0}{k} - T_{amb}\right) e^{-\frac{kt}{c_0 m}} + \frac{p_0}{k} + T_{amb} \tag{45}
$$

令 $c_0 = \frac{C}{m}$，　$k = \frac{1}{R}$，　$p_0 = P$，可得到与 3.3 节中式(11)形式完全一样的结果，即：

$$
T = PR + T_{amb} + \left(T_0 - PR - T_{amb}\right) e^{-\frac{t}{RC}} \tag{46}
$$

LI Xiang, born in 1989, Ph. D. candidate. His research interests include cloud computing, energy efficiency of data center, cloud modeling and simulation.

Jiang Xiao-Hong, born in 1966, Ph. D., associate professor. Her research interests include computer architecture, distributed systems, cloud computing, etc.

WU Zhao-Hui, born in 1966, Ph. D., professor. His research interests include service science and grid computing, embedded systems, ubiquitous computing, etc.

XE KE Jiang, born in 1986, Ph. D. candidate. His research interests include virtualization and cloud computing, performance evaluation and modeling.

Background

The high energy consumption of data center is a serious problem to be solved. Especially as the development of cloud computing, more resources are centralized to the clouds. Constructing green data centers and achieving power cost and carbon footprint reduction become research hotspots in recent years. A lot of work dedicated to thermal management and thermal balance has been done. This paper surveys the latest research result of thermal management for green data centers from the perspective of monitoring, modeling, management and evaluation.

This work is supported by National High Technology Research and Development Program (863 Program) of China (No. 2011AA01A207) and National Natural Science Foundation of China (No. 61272128). These projects aim to study the energy-efficient cloud computing technology, theory and methods. Our group has been working on this project for several years and published many related papers on this topic. This paper is a survey of the issues of thermal management for green data centers.