文章编号:1671-4598(2023)11-0007-09 DOI:10.16526/j.cnki.11—4762/tp.2023.11.002 中图分类号:TP301 文献标识码:A

面向数据中心的服务器能耗模型综述

王东清，李道童，彭继阳，叶丰华，张炳会
（浪潮电子信息产业股份有限公司，北京 100085）

摘要：伴随着云计算技术的快速发展，数据中心的服务器能耗日益激增，带来了严重的经济和环境问题，降低数据中心能耗，对缩减数据中心运营成本、实现全球“双碳”战略目标具有重要意义；因此，不同层面的服务器能耗模型构建和预估成为了近年来研究的热点；据此，从硬件、软件层面系统地总结了服务器能耗模型的相关工作；在硬件层面，对服务器的整体能耗按加法模型、基于系统利用率模型和其他模型分类；同时，还总结了服务器部件粒度的能耗模型，涵盖 CPU、内存、磁盘和网络接口；在软件层面，按机器学习的类别将服务器能耗模型归纳为监督学习、非监督学习、强化学习；此外，还比较了不同能耗模型的优缺点、适用场景，展望了能耗模型的未来研究方向。

关键词：云计算；数据中心；能耗模型；监督学习；非监督学习；强化学习

A Survey of Server Energy Consumption Models in Data Center

WANG Dongqing, LI Daotong, PENG Jiyang, YE Fenghua, ZHANG Binghui
(Inspur Electronic Information Industry Co., Ltd., Beijing 100085, China)

Abstract: With the rapid development of cloud computing, the increasing demand for server energy consumption in data centers leads to crucial economic and environmental issues. Reducing the data center energy consumption is of great significance to save the operating cost of data centers and realize the global “double-carbon” strategic goal. Therefore, the construction and prediction of server energy consumption models at different levels become a research hotspots in recent years. Accordingly, the relevant work of server consumption models is systematically summarized from two levels of hardware and software. At the hardware level, the overall energy consumption models of the cloud server are classified from the additive models, models based on system utilization rate and other models. Meanwhile, the energy consumption models of the server components are also summarized, including the CPU, memory, disk and network interface. At the software level, the server energy consumption models are summarized according to the category of machine learning, such as supervised learning, unsupervised learning and reinforcement learning. Additionally, the advantages, shortcoming and suitable scenarios of different consumption models are also compared, which prospects the future research directions of consumption models.

Keywords: cloud computing; data centers; energy consumption models; supervised learning; unsupervised learning; reinforcement learning

0 引言

云计算及其应用（如 5G、大数据、互联网）的兴起，推动了数据中心的蓬勃发展。数据中心作为支撑云服务的基础设施，有着大量的服务器及硬件配套设施，需要耗费巨大的电能保障运行，这种能耗随着数据中心的规模扩大在逐年剧增[1]。一个普通的数据中心用电量相当于 25 000 个家庭，并以每 5 年翻倍的速度增加[2]。据统计，2017 年全球数据中心能耗总量占全球能耗总量 3%[3]；到 2030 年，该占比预估会增至 8%左右[4]。能耗正成为数据中心运营的主要成本，甚至超出购买设备的初始成本[5]。另一方面，为数据中心供电所需的能耗会导致温室气体的排放，占全球排放量的 2%[6]，违背全球“双碳”战略目标。

相比数据中心硬件设备本身的电量消耗，服务器资源利用率低下、负载和功率不匹配造成的能效低下也是能耗问题的重要原因。举例来说，运营商通常采用冗余资源部署策略，保障高性能、高可靠的服务，导致主机的 CPU 正常利用率仅为 10%～50%[7]。此外，即使在空闲模式下运行，服务器也会消耗大量的电量，功率可达到峰值的 70%[8]。

综上所述，维持数据中心大量基础设施运行所需的电量消耗，会带来二氧化碳大量排放、投入成本高等负面影响，降低数据中心能耗对提供“绿色云计算”、实现“双碳”战略目标具有重要意义。数据中心能耗主要源自两部分：IT 设备（服务器、网络设备）和基础设施（冷却、电力调节系统）。据统计[9-12]，各模块能耗占比为：服务器 56%、网络设备 5%、冷却系统 30%、电力调节系统 8%、其他（如照明）1%。因此，数量庞大的服务器是数据中心能耗最大的组件，其他组件的能耗也与服务器有关，降低

收稿日期:2023-07-22； **修回日期**:2023-08-29。
基金项目:山东省基金项目(2019LZH006)。
作者简介:王东清(1989-),男,研究员,博士。
引用格式:王东清,李道童,彭继阳,等.面向数据中心的服务器能耗模型综述[J].计算机测量与控制,2023,31(11):7-15.

其冗余的能耗是解决数据中心能耗的重要且有效的途经。

为提高服务器资源利用率和整体能效，能耗建模和预估一直是学术界和产业界关注的热点。王继业等[13]从能效模型与能效算法层面总结了数据中心服务器系统与网络系统的节能研究进展。余潇潇等[14]对数据中心主要负荷部分能耗模型进行了回顾和分类，并对能量调节的可行性和工作方式进行了总结。罗亮等[15]采用多元线性回归和非线性回归的数学方法，分析总结了不同参数和方法对服务器能耗建模的影响。JIN[16]根据计算公式等因素将现有功耗模型分类，通过比较发现，多项式模型和线性回归模型在精度方面表现更好。STANLY[17]讨论了传统的粒子群优化算法和遗传算法在数据中心能耗中的应用。上述文献对数据中心能耗方面的研究进行了较好的归纳总结，但缺少一定的系统性，尤其是近年来机器学习、深度学习模型在能耗预估方面的应用。因此，从硬件、软件层面系统地总结了服务器各类能耗模型的相关工作，并比较了各自的优缺点、适用场景。

1 服务器能耗模型

服务器作为数据中心的核心组件，对其构建精准、有效的能耗模型在供电系统设计、节能提效、新设备采购等方面有重要指导意义。

为此，归纳总结服务器及其部件（如图 1 所示）在能耗建模方面的研究进展，具体如下：1）在硬件层面，总结了服务器整体的能耗模型分类，即加法模型、基于系统利用率模型、其他模型；此外，也对服务器的构成部件对能耗模型分类总结；2）在软件层面，将服务器能耗模型研究成果分为监督学习、非监督学习、强化学习。本节将根据能耗模型的计算方式不同，介绍加法模型、基于系统利用率模型和其他模型。在具体模型中，会涉及服务器的部件（如 CPU、内存、磁盘、网络接口）能耗模型，后续章节会有详细描述。

1.1 加法模型

该类方法将服务器整体能耗以各部件能耗求和方式计算，将根据模型计算复杂度由简到繁介绍。ROY 等人[18]提出了一种简单的服务器能耗模型，表示如下：

$$E_{\text{server}} = E_{\text{cpu}} + E_{\text{mem}} \tag{1}$$

其中：E_{cpu}和 E_{mem}分别表示处理器和内存的能耗模型。为了更精准描述服务器能耗，研究者增加服务器磁盘、网络接口等部件的能耗模块[19-20]：

$$E_{\text{server}} = E_{\text{cpu}} + E_{\text{mem}} + E_{\text{disk}} + E_{\text{NIC}} \tag{2}$$

研究人员[20]根据能耗、功耗和时间三者间的关系，将服务器能耗模型表示为部件平均功耗与运行时间乘积，再对各部件的求和：

$$E_{\text{server}} = \bar{P}_{\text{comp}} T_{\text{comp}} + \bar{P}_{\text{NIC}} T_{\text{comp}} + \bar{P}_{\text{network}} T_{\text{network}} \tag{3}$$

$\bar{P}_{\text{comp}}$表示 CPU 和内存平均功耗，T_{comp}是平均计算时间，$\bar{P}_{\text{NIC}}$为网卡平均功耗，$\bar{P}_{\text{network}}$为网络设备平均功耗，T_{network}是网络设备运行时间。

文献［21］采用了部件信息更丰富、关系相对更复杂的能耗计算形式：

$$E_{\text{server}} = c_0(E_{\text{cpu}} + E_{\text{mem}}) + c_1 E_{\text{em}} + c_2 E_{\text{board}} + c_3 E_{\text{hdd}} \tag{4}$$

$c_0,\cdots,c_3$ 表示权重系数，需通过线性回归模型学习，这些系数在特定的服务器中是不变的。E_{cpu}，E_{mem}，E_{em}，E_{board}，E_{hdd}分别表示处理器、内存、机电、主板和硬盘的能耗。

另一种加法模型是将虚拟机（virtual machine，VM）作为部件选项[22]：

$$P_{\text{server}} = P_{\text{baseline}} + \sum_{i=1}^{n} P_{\text{vm}}(i) \tag{5}$$

P_{baseline}为基线功耗，是个经验值。P_{vm}表示虚拟机功耗，n 为服务器拥有的虚拟机数量。根据[25]研究，P_{vm}可表征为处理器、内存、I/O 等部件利用率或吞吐量的加权和：

$$P_{\text{vm}} = \alpha_1 U_{\text{cpu}} + \alpha_2 U_{\text{mem}} + \alpha_3 U_{\text{io}} + e \tag{6}$$

公式（5）可进一步扩展为[22-23]：

$$P_{\text{server}} = P_{\text{baseline}} + \alpha_1 \sum_{i=1}^{n} U_{\text{cpu}}(i) + \alpha_2 \sum_{i=1}^{n} U_{\text{mem}}(i) + \alpha_3 \sum_{i=1}^{n} U_{\text{io}}(i) + ne \tag{7}$$

1.2 基于系统利用率模型

该类方法将服务器整体能耗以各部件利用率求和方式计算，考虑到 CPU 是服务器中的能耗主要来源，该类模型的研究主要围绕 CPU 展开。

最早的基于系统利用率的服务器能耗模型是由 ELNOZAHY 等人[24]提出：

$$P_f = c_0 + \alpha^2 A C f^3 \tag{8}$$

A 表示切换活动（如单位时钟内切换的数量），C 为电容容量，f 为时钟频率，α 表示常数系统，c_0 表示除 CPU 以

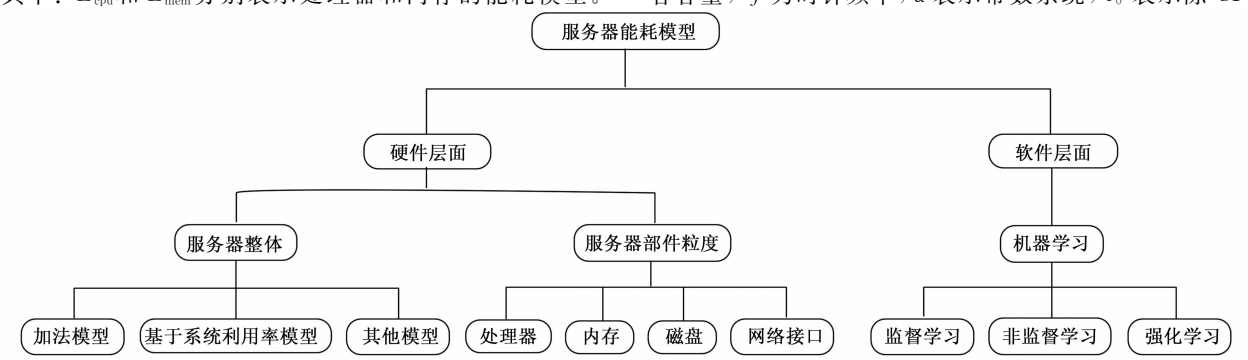

图 1 服务器能耗模型分类标准

外的部件功耗。

部分研究[26-27]假设服务器功耗与处理器利用率成正相关，提出了一种对后续研究有重大影响力的能耗模型：

$$P_u = (P_{\max} - P_{\text{idle}})\mu + P_{\text{idle}} \tag{9}$$

P_{idle}和 $P_{\max}$分别表示服务器空载和满负载下的平均功耗，μ 属于［0，1］区间的加权系数，代表 CPU 的利用率。

类似地，FAN 等人[28-31]提出一种非线性的系统利用率能耗模型：

$$P_u = (P_{\max} - P_{\text{idle}})(2\mu - \mu^r) + P_{\text{idle}} \tag{10}$$

其中，参数 r 是个人工校准参数，用以最小化模型拟合误差。文献评估上百个服务器能耗，结果表明模型（10）比模型（9）效果更好。

文献［31］在模型（9）的基础上，给出了一种更复杂的能耗模型：

$$P_x(t) = (P_{x\_full} - P_{x\_idle})\alpha_x U_x(t)^{\beta_x} + P_{idle} \tag{11}$$

P_{x_full} 和 P_{x_idle} 分别表示服务器在满负载和空负载下的能耗，α_x 和 β_x 为和服务器相关的超参。$U_x(t)$ 为 t 时刻的 CPU 利用率。

YAO 等人[32]描述了另一种能耗模型：

$$P = \frac{b_i(t)^{\alpha}}{A} + P_{\text{idle}} \tag{12}$$

其中：A，P_{idle}和 α 均为服务器相关的常数。P_{idle}为服务器空闲负载时的平均功耗。$b_i(t)$ 对应 t 时刻的服务器利用率。

1.3　其他模型

大多数的服务器能耗研究都属于加法、基于系统利用率模型，本节将介绍其他类型的能耗模型。

LEFURGY 等人[33]观察服务器的运行状态发现，当系统的性能表现变化时，服务器的功耗也立刻改变。因此，研究人员认为特定状态下的工作负载，服务器的功耗由性能状态决定，与过去的控制周期能耗无关。据此，提出以下能耗模型：

$$p(k) = At(k) + B \tag{13}$$

其中：A 和 B 为系统相关参数，$p(k)$ 为服务器在第 k 个控制周期下的能耗，$t(k)$ 对应 CPU 在第 k 个控制周期的性能状态。

此外，有研究[34]以标准的 M/M/1 排队理论为基础，假设服务器能耗与 CPU 使用率成线性关系，捕捉服务器中的请求处理行为，能耗模型公式如下：

$$P(\gamma) = \frac{\gamma}{\mu}(P_{\text{cpu}} + P_{\text{other}}) + (1 - \frac{\gamma}{\mu})P_{\text{idle}} \tag{14}$$

P_{cpu}和 P_{other}分别表示处理器和其他系统的能耗。$\frac{1}{\mu}$ 和 $\frac{1}{\gamma}$ 分别表示请求时间间隔、服务时间为指数分数时对应的参数。

2　服务器部件粒度能耗模型

由于服务器能耗在数据中心的能耗占比中不可忽略，拆分其组成部件、建立每个部件的能耗模型，对提升服务器的资源利用率、降低能耗不可或缺。图 2 给出服务器各部件的能耗占比[35-36,45]，其中处理器能耗占比最大，其他依次为内存、其他、磁盘和网络接口，本节将重点介绍这些部件的能耗模型。

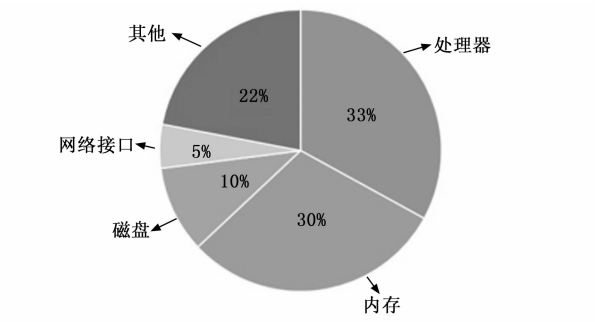

图 2　服务器部件能耗占比

2.1　处理器能耗模型

2.1.1　单核 CPU 能耗模型

鉴于多核 CPU 能耗模型衍生自单核，调研单核 CPU 能耗模型是很有必要的。单核 CPU 能耗模型分为加法模型和基于性能计数器能耗模型两种。

一种典型的 CPU 加法能耗模型如下所示[37]：

$$P_{\text{cpu}} = P_d + P_s + P_o \tag{15}$$

其中：动态能耗 $P_d = ACV^2 f$，A、C 和 f 含义与公式(8)一致，V 表示供电电压。静态功耗 P_s 来自泄电流功耗，又细分为亚阈值漏流和栅漏流功耗，P_o 表示常开功耗。

通过聚合 CPU 的架构部件功耗，BERTRAN 等人[38]提出如下能耗模型：

$$P_{\text{total}} = P_{\text{static}} + (\sum_{i=1}^{n} A_i P_i) \tag{16}$$

$A_i P_i$ 表示第 i 个架构部件的动态功耗，A_i 为对应的活动率，P_i 为对应的加权系数，n 为 CPU 架构部件总量，P_{static} 为所有部件的静态功耗。

如今 CPU 提供了丰富的性能计数器，CHEN 等人[39]指出，主要有 5 种性能计数器对精准预测 CPU 能耗是有效的：

$$P = c_0 + c_1 L_{1ps} + c_2 L_{2ps} + c_3 L_{2mps} + c_4 F_{pps} + c_5 B_{rps} + c_6 f^{1.5} \tag{17}$$

L_{1ps} 为单位时间内 L_1 数据缓存引用数，L_{2ps} 为单位时间内 L_2 缓存引用数，L_{2mps} 为单位时间内 L_2 缓存丢失数，F_{pps} 为单位时间内的浮点计算数，B_{rps} 为单位时间内过期的分支指令数，f 为 CPU 频率，c_0，…，c_6 为权重系数，c_0 为系统空闲和泄电流功耗。

2.1.2　多核 CPU 能耗模型

目前，数据中心服务器配置的基本为多核 CPU，不同线程对内核使用情况的差异，会产生不同水平的 CPU 能耗，决定了建立 CPU 核心级的能耗模型是很有必要的。多核 CPU 能耗模型的研究分为队列模型和加法模型两种。

文献［40］将多核 CPU 等价为具有多个服务器的 $M/M/m$ 队列系统，采用两种内核加速模式：a）空闲加速模式（无任务运行）；b）常数加速模式（无论是否有任务运行，所有内核以固定速度运行），在常数加速模式下，CPU 的能

耗表达式如下：

$$P_{\text{total}} = \sum_{j=1}^{n} P_c(j) \tag{18}$$

P_{total}是 n 个内核的总功耗，$P_c(j)$ 对应第 i 个内核功耗，单个内核的功耗可采用 2.1.1 中的建模方法。在此基础上，通过对不同加速模式下的单核模式改进，可提升队列能耗模型的预估精度。

另一种多核 CPU 功耗模型是计算线程功耗的总和，如文献［41］将 CPU 功耗理解为多核的负载总和：

$$P = (P_{\text{idle}} + CP_t)T \tag{19}$$

C 表示并发负载量，P_t 为单线程能耗，T 为完成负载的执行时间。CP_t 表示处理器所有线程完成负载时的总功耗。

2.1.3 GPU 能耗模型

人工智能的快速发展，带来 GPU 的需求量日益激增，GPU 服务器也成为了数据中心的标配。本节将介绍基于性能计数器和加法两种能耗模型。

Song 等人借助统计学习方法，根据来自硬件性能计数器的数据分析，构建 GPU 的性能能耗模型，所提的方法可以直接从数据层面分析，无需深入理解 GPU 底层运行逻辑。

作为加法模型，文献［42］将 GPU 的功耗以为各部件求和方式计算：

$$P = P_{\text{fpu}} + P_{\text{alu}} + P_{\text{mem}} + P_{\text{ot}} \tag{20}$$

P_{fpu}，P_{alu}，P_{mem}，P_{ot}分别对应浮点运算单元、逻辑计算单元、内存和其他部分功耗。类似的，文献［43］通过增加更多的组件等方式对 GPU 的能耗来源划分。

2.2 内存能耗模型

内存的构成部件存在明显的层级结构，DRAM 是其主要的构成部件，DRAM 大容量和高带宽特性决定了内存在服务器能耗中的高占比。

作为加法模型的经典用法，动态和静态能耗也被用在内存能耗模型中。LIN[44]采用了一种简单的数学公式表达 DRAM 能耗：

$$P = P_{\text{static}} + c_1\mu_{\text{read}} + c_2\mu_{\text{write}} \tag{21}$$

c_1，c_2 为读写操作下的功耗参数，P_{static}为 DRAM 静态功耗。$c_1\mu_{\text{read}} + c_2\mu_{\text{write}}$ 为动态功耗，μ_{read} 和 μ_{write} 分别为读写吞吐率。

另一种基于部件的加法功耗模型为[46]：

$$E = E_{\text{icache}} + E_{\text{dcache}} + E_{\text{bus}} + E_{\text{pads}} + E_{\text{mm}} \tag{22}$$

E_{icache}和 E_{dcache}分别表示指令高速缓存、数据高速缓存能耗；E_{bus}表示在数据和 icache/dcache 间对应的地址总线和数据总线能耗；E_{pads}表示 I/O 版和外部总线到内存缓存的能耗；E_{mm}表示内存缓存能耗。

2.3 磁盘能耗模型

磁盘可分为固态硬盘和机械硬盘。其中，固态硬盘因其容量小、价格昂贵、寿命有限等缺陷，暂未广泛使用。目前，数据中心还是以机械硬盘作为主要存储介质。因此，本节主要介绍机械硬盘的能耗模型。

简单的磁盘能耗模型可根据空闲、运行负载两种状态求和预估[47]，其中，运行负载能耗主要来自磁臂头和盘片旋转的移动。相比之下，一些研究指出[48]，磁盘运行功耗与 I/O 请求率相关，空闲功耗占峰值功耗 2/3，两者并非简单的线性关系，对应能耗模型为：

$$P = P_{\text{idle}} + a_1 q + a_2 q^2 \tag{23}$$

a_1，a_2 为待定参数，q 表示磁盘单位时间内的 I/O 请求数。

2.4 网络接口能耗模型

有研究指出[7]，数据中心的平均使用率比较低。在系统处于低利用下，网络接口依然处于满负载运行状态，造成大量的能耗资源浪费。因此，构建网络接口能耗模型，动态调整网络接口运行状态，可进一步降低数据中心能耗。

当网络接口处于空闲或备用状态，P_{idle}表示对应状态功耗，t_{idle}表示对应 d 状态时间。当网络接口接收或传输数据包时，即为激活状态，P_{active}表示对应状态功耗，t_{active}表示该状态运行时间 。E_s 为不同状态切换所需能耗，n_s 为状态切换次数，能耗模型表示如下[49]：

$$E = P_{\text{idle}}t_{\text{idle}} + P_{\text{active}}t_{\text{active}} + E_s n_s \tag{24}$$

3 基于机器学习的能耗模型

能耗模型作为数据中心服务器实现系统性监控、预测的重要基础，建模方法的优化、预测精度的提升一直是数据中心能耗研究者追求的目标。近年来，随着机器学习技术的快速发展，也成为了能耗研究者的研究热点[77]，各种经典的机器学习算法和数据挖掘方法被广泛应用于能耗建模、预测。根据统计学基础，机器学习方法可分为三类：监督学习方法、无监督学习方法和强化学习方法。本节将依据机器学习的分类对能耗模型文献进行总结。

3.1 监督学习能耗模型

监督学习是从有标签的数据学习模型，并对新的输入数据预测结果。在能耗建模、预测问题中，输入可以是服务器中的任何数据表征，输出为服务器的能耗大小。典型的监督学习模型包括线性回归、支持向量机、决策树、神经网络、集成学习等。

1）单因素线性回归：文献［50］根据服务器能耗与 CPU 利用率呈二次关系，构建了能耗模型（25）。类似地，模型（26）采用 r 次多项式[52]，避免线性回归中的过拟合问题。2013 年，文献［51］对 SPECpower 打榜应用中 7 种不同服务器的能耗数据分析发现，扩展线性回归中的多项式阶数可以更精准的预估能耗，如公式（27）。

$$P = \alpha + \beta_1 u_{\text{cpu}} + \beta_2 u_{\text{cpu}}^2 \tag{25}$$

$$P = \alpha + \beta_1 u_{\text{cpu}} + \beta_2 u_{\text{cpu}}^r \tag{26}$$

$$P = \alpha + \beta_1 u_{\text{cpu}} + \beta_2 u_{\text{cpu}}^2 + \beta_3 u_{\text{cpu}}^3 \tag{27}$$

其中：α，β_1，β_2，β_3 为模型参数，可根据优化算法获得，r 为超参，由验证数据集选定。

2）多因素线性回归：增加更多的特征元素是提升线性回归模型精度的常用手段，但也会随着特征元素的增加引入过拟合问题，Lasso 回归模型以引入正则项的方式平衡预

测精度和训练过拟合问题。公式（27）表示最小化 Lasso 回归模型的目标函数，y 为模型拟合的目标值，X 为输入的特征元素，W 为模型系数，α 为正则项系数。

$$\min \| y - XW \|_2^2 + \alpha \| W \|_1 \tag{28}$$

借鉴上述思路，能耗研究者也更多的关注模型的丰富特征输入[53-54]。相比文献［50－52］，研究[53]增加了内存利用率来建立与服务器的能耗关系。作者提出使用 CPU 和内存的三阶多项式来作为 Lasso 模型的特征输入，即 $X=\{\text{cpu}, \text{cpu}^1, \text{cpu}^2, \text{cpu}^3, \text{mem}, \text{mem}^1, \text{mem}^2, \text{mem}^3\}$。此外，文中也提出了一种指数形式的三阶多项式，即 $X=\{e^{\text{cpu}}, e^{\text{cpu}^1}, e^{\text{cpu}^2}, e^{\text{cpu}^3}, e^{\text{mem}}, e^{\text{mem}^1}, e^{\text{mem}^2}, e^{\text{mem}^3}\}$ 进一步提升模型精度，［54］采用了 30 个不同的硬件资源利用率特征挖掘与服务器能耗的线性关系。

3）支持向量机：文献［54］发现 CPU 和内存的利用率是相互影响的，通过线性模型预测服务器的能耗会有较大精度损失。因此，他们提出一种基于支持向量机的回归（support vector regression，SVR）模型，如公式（28），旨在寻找一个线性超平面，用以拟合多维输入特征与能耗值之间的非线性关系：

$$f(x) = w\varphi(x) + b \tag{29}$$

$$\min_{w,b,\zeta} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} (\zeta_i + \zeta_i^*) \tag{30}$$

$\varphi(x)$ 为 CPU 和内存利用率在核函数中的映射空间。参数 w 和 b 可通过最小化公式（29）的目标函数计算。C 表示错误惩罚因子，ζ_i, ζ_i^* 分别为松弛变量的上下界。LUO 等人[55]也建立了 SVR 模型用来预测服务器的能耗，并比较了 Lasso 回归模型、逐步回归模型在 SPECcpu _ 2006 打榜数据的表现，证明了 SVR 模型在预测精度方面相比其他模型的优势。此外，为解决 SVR 模型对输入特征质量敏感问题，YANG 等人[56]采用 PCA 算法对预处理特征，提升 SVR 模型能耗预测精度。

4）集成学习：简而言之，通过学习多个基学习器，并按某种策略组合输出目标的过程。由于采用群体决策的过程，相比单个模型，可缓解模型训练中的过拟合或欠拟合问题。根据基学习器之间是否存在依赖关系，集成学习可分为并行和串行两种训练模式，并行和串行集成学习的典型代表分别提升算法和随机森林。在服务器能耗建模方面，随机森林或装袋算法用的相对较多。例如，HARTON 等人[57]在 CPU 利用率处于动态变化场景下，提出一种基于随机森林算法的服务器能耗预测框架。LIN 等人[60]以服务器多个部件的信息作为特征，建立以 CART 为基学习器的随机森林能耗模型，实验结果表明，所提模型具有高精度、高鲁棒性等优势。此外，诸如 GBDT[59]、xgboost[58]等提升方法在预测任务有更好的表现，也可以被用来建立更高效的能耗模型。

5）神经网络：人工神经网络（ANN，artificial neural network）因具有自主学习功能、泛化能力强、容错性高等优势，在多个领域得到广泛应用，尤其是深度学习进入视野之后。特别地，在能耗预估方向，ANN 也成为研究热点。2016 年，LI 等人[61]使用了细粒度和粗粒度两种深度学习模型，预测不同时间尺度下的能耗。递归自编码器作为细粒度模型被用以预取短周期内的能耗预测，相应地，自编码器作为粗粒度模型通过对短周期历史数据编码，预测长周期的能耗。2017 年，LIU 等人[62]提出一种基于神经网络的优化方法降低数据中心的能耗，具体地，利用反向传播神经网络（BPNN，back propagation neural network）和长短期记忆（LSTM，long short-term memory）神经网络预测 Google 数据中心在未来 5 min 内的能耗。LIN 等人[63]深入分析 CPU、内存和磁盘的能耗特性，从中选择一组代表服务器能耗特性的部件计数器信息作为模型输入，评估 BPNN、ENN（ENN，elman neural network，一种简单的循环神经网络）和 LSTM 网络三种模型在不同负载下的预测效果，实验结果表明，相比多因素回归、SVR 模型，所提的三种模型在预测实时能耗方面效果更优。此外，神经网络也被用于数据中心负载和资源预测任务[64-65]，尽管目标不同，这些研究的建模方法也适用于服务器能耗预测问题，本质上它们都是一种基于时间序列的预测任务。

3.2　非监督学习能耗模型

不同于监督学习，监督学习是从不含标签的样本中训练模型参数的。聚类算法（如 K-means、层次聚类）和高斯混合模型（GMM，gaussian mixture models）是两种典型的非监督学习技术。在能耗模型中，模型的输出本质上相当于从预估的能耗分布中抽取样本。因此，研究者会选择 GMM 或 K-means 建立能耗模型。

GMM 是扩展的高斯混合模型，通过组合多个高斯分布表示真实的数据分布。每个单独的高斯分布代表一个隐藏变量，并配有相应的权重系数，计算所有隐藏变量的加权系数得到最终的输出结果，如真实的能耗值，GMM 的数学表达：

$$p(y \mid \theta) = \sum_{k=1}^{K} \alpha_k \varphi(y \mid \theta_k) \tag{31}$$

其中：α_k 对应第 k 个隐藏变量的权重系数 $\sum_{k=1}^{K} \alpha_k = 1$，$\varphi(y \mid \theta_k)$ 对应第 k 个隐藏变量在观察变量 y 下的高斯概率密度值，$\theta_k = (\mu_k, \sigma_k^2)$ 为第 k 个隐藏变量的高斯分布参数。理论上，GMM 在聚合足够多的高斯模型，并配以正确的权重系数，混合模型会正确的反映真实的能耗数据分布。因事先不清楚观察样本属于哪个隐藏变量分布，如能耗建模中不清楚影响能耗的行为有哪些，GMM 中的参数集 $(\alpha_k, \mu_k, \sigma_k^2)$ 无法采用单个高斯模型中最大似然估计法求解。最大期望算法（EM，expectation maximization）作为一种迭代算法，可用于求解含有隐藏变量的参数估计，如 GMM、K-means 算法。以 GMM 能耗建模为例，EM 优化算法可分为 3 步：1）初始化隐藏变量权重系数及对应的高斯分布参数；2）根据已有的高斯分布参数，计算能耗观察样本属于每个高斯模型的概率；3）基于能耗观察样本属于每个高斯模型的概率，计算新一轮高斯模型的参数，重复过程 2）和 3）直到

参数收敛。DHIMAN[66]提出一种基于 GMM 的能耗预估模型，文献根据 CPU 利用率对观察数据分组，并训练 GMM 参数。在能耗预测时，对 GMM 中不同的高斯模型计算距离，选择最近的高斯模型来预测能耗。2017 年，ZHU 等人[67]采用 GMM 模型挖掘服务器能耗与 CPU、内存利用率间的关系。此外，作者还分析不同特征配置下 GMM 能耗预估效果。

MYDHILI[68]提出一种多步并行的 K-means 算法协调数据中心的能耗资源。更进一步，文献 [69] 融合 K-means 和帝国竞争算法（imperialist competition algorithm，ICA）提升数据中心的资源利用率、降低能耗。

3.3 强化学习能耗模型

强化学习是一种不同于监督学习、非监督学习的算法派系，其基本原理为智能体和环境在交互的过程中学习，获取环境反馈的奖励，不断调优自身的策略。组成强化学习的 5 个基本元素为：1）环境状态集合 S（State）：描述智能体在 t 时刻所处环境情况 s_t，其中 $s_t \in S$，S 为所有状态的集合；2）智能体的动作集合 A（Action）：描述智能体在 t 时间采取的动作 a_t，其中 $a_t \in A$，A 为所有动作的集合；3）环境的奖励 R（Reward）：表示智能体 t 时刻在状态 s_t 执行动作 a_t 的瞬时奖励 R_{t+1}，用以衡量当前行为对最终目标的贡献；4）智能体的策略 π：描述一个函数，输入为 s_t，输出动作 a_t 在状态下被执行的概率，即 $\pi = p(a_t \mid s_t)$；5）价值函数：瞬时激励是对单次行为或环境状态的贡献评估，强化学习的目标是期望长期最优回报，这种基于长期回报的函数被成为价值函数，基于评价对象不同又分为状态价值函数和动作状态价值函数，对应表达式分别为（32）和（33）：

$$v(s_t) = E(R_{t+1} + \gamma v(s_{t+1}) \mid s_t) \tag{32}$$

$$q(s_t, a_t) = E(R_{t+1} + \gamma q(s_{t+1}, a_{t+1}) \mid s_t, a_t) \tag{33}$$

其中：γ 为折扣因子，计算长期回报。强化学习训练的具体过程为：智能体获取环境状态信息 s_t，根据策略 π 选择执行动作 a_t；执行动作后的环境进入新的状态 s_{t+1}，同时，给出动作对应的瞬时激励 R_{t+1}；智能体根据得到的激励优化策略参数，在新的策略参数下，对状态 s_{t+1} 执行新的动作选择策略；上述过程不断重复，直至价值函数收敛。

在实际生产中，强化学习被成功应用于各个领域，如自动驾驶[70]、游戏[71]、机器人控制[72]。在数据中心领域，强化学习也被用于服务器资源调度和能耗管理中，降低服务器中冗余的能源消耗[73-76]。如前文所述，服务器的资源利用率相对较低，通过整合、降低主机的无效资源，是有效的节能手段。据此，研究者[73]采用强化学习动态调整集群中运行的主机数量。LIN[74]提出一种基于强化学习的服务器能耗管理方法，可不用考虑作业达到或作业服务过程的限制，经评估公开的谷歌服务器数据，强化学习可在作业响应时间内有效降低能耗。2017 年，LIU 等人[75]以强化学习为基础，提出一种面向数据中心的服务器资源调度和能耗管理框架，该框架主要由负责服务器虚拟机资源调度分配的全局层和负责本地能耗管理的本地层组成。

4 服务器能耗模型发展方向展望

从硬件、软件层面总结了各种能耗模型的原理及应用。然而，因组成服务器的硬件架构、虚拟机、上层应用等复杂多样，建立好的能耗模型仍然需要研究者付出更多的努力。其中，能耗模型的有效性是实际应用中关注的重点。对比软硬件建模方法，一些硬件系统的方法在预估准确性上更高，归因于硬件方法从服务器硬件架构出发，有更明确的物理意义。以公式（1）～（5）中的模型为例，通过提高服务器的 CPU、内存、硬盘和网络接口等部件粒度的能耗精度，对进一步提高服务器整体能耗预估准确性会有积极意义。在基于系统利用率的能耗模型研究者，CPU 被认为是服务器能耗的主要来源，对于以 GPU 或 FPGA 为架构核心的服务器，CPU 的能耗模型可能不适用，可以考虑其他的能耗模型，如构建 GPU/FPGA、内存、磁盘等部件与能耗的关系。尽管如此，这些基于硬件系统的低层级能耗建模方法不能适用所有场景，如更上层的操作系统、虚拟机、应用软件等。一方面是很多数据中心的硬件基础设施会随着技术迭代、客户需求等因素不断更新换代，另一方面，更高层的能源消耗，需依托机器学习算法才能更有效的挖掘各层级与能耗的复杂关系。

训练数据的采集和模型参数的选择是机器学习能效建模准确性的基础保证。尽管通过采集虚拟机或容器与能耗相关的间接信息已被证明可用来建立机器学习能耗模型，通过外部设备直接采集能耗信息更有助于能耗模型的准确性。此外，非监督学习因其对隐藏变量（如单个虚拟机或容器的能耗分布可看作隐藏变量）的假设，为建立服务器整体能耗模型提供理论基础。支持向量机在小数据集的鲁棒性，也可作为备选方案。此外，强化学习是在和服务器系统交互过程中获取奖励反馈，不断学习调整策略的，无需事先理解服务器的能耗分布，相比其他机器学习方法，可扩展性更强。另一方面，机器学习算法建立模型时的参数选择值得深入探讨的问题。以 K-means 方法为例，聚类中心的个数、初始化都会影响能耗模型的最终效果。同样地，SVM 核函数的选择会导致模型的效果、计算复杂度差异。神经网络算法的设计需考虑隐藏单元个数、激活函数、正则化方法等因素的影响，避免过拟合、泛化能力差等现象。

尽管关于服务器能耗模型取得一定进展，如何在实际中建立更好的能耗模型仍需进一步的探索，具体如下：1）大多数研究以能耗作为指标评估服务器能效，探索一种融合能耗、可靠性、响应时间的多目标能耗调度模型可为用户提供更高质量的服务；2）机器学习、深度学习模型需要进行参数选择，易导致过拟合、欠拟合等问题，需要丰富的模型参数调优经验，费时费力，采用自动化结构搜索的方法可改善开发效率、提升预估准确性，如神经网络结构搜索技术；3）如引言所述，服务器和冷却系统是数据中心能耗主要来源。研究表明，当服务器的使用率越高时，产生的热量越多，冷却系统也需要耗费更多的电力为服务器

降温。因此，在能耗模型设计时，联合优化服务器和冷却系统的标注信息，对数据中心整体能耗降低具有重要的意义；4）跨区域的数据中心正在成为发展趋势，为节能目标提供新的方向。例如，采用地理空间负载平衡技术，在模型中融入请求用户与物理节点距离、能源类型、能源价格、实时温度和整体负载等地域差异性信息，探索一种跨区域数据中心节点间资源任务调度和能效优化模型。

5　结束语

随着互联网、5G、大数据等技术对算力需求的增加，云计算的规模也不断扩大，数据中心的服务器能耗问题日益凸显，改进服务器的能源管理策略成为数据中心运营商的工作重点，能耗模型的构建和预估在其中扮演重要的角色。因此，系统性地从软硬层面归纳总结了服务器能耗模型的相关工作，并对每个层面的能耗模型进行分类，对降低数据中心能耗具有重要指导意义。此外，还比较了不同能耗模型的优缺点、适用场景，对能耗模型的未来研究方向进行了展望。

参考文献：

[1] DAYARATHNA M, WEN Y, FAN R. Data center energy consumption modeling: A survey [J]. IEEE Communications Surveys & Tutorials, 2015, 18 (1): 732 - 794.

[2] KAPLAN J M, FORREST W, KINDLER N. Revolutionizing data center energy efficiency [J]. McKinsey & Company, 2008: 1 - 13.

[3] DANILAK R. Why energy is a big and rapidly growing problem for data centers [J]. Forbes, 2017, 15: 12 - 17.

[4] JONES N. How to stop data centres from gobbling up the world's electricity [J]. Nature, 2018, 561 (7722): 163 - 166.

[5] RIVOIRE S, SHAH M A, RANGANATHAN P, et al. Models and metrics to enable energy-efficiency optimizations [J]. Computer, 2007, 40 (12): 39 - 48.

[6] BUYYA R, BELOGLAZOV A, ABAWAJY J. Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges [J]. arXiv preprint arXiv: 1006. 0308, 2010, 12 (4): 6 - 17.

[7] BARROSO L A, H LZLE U. The case for energy-proportional computing [J]. Computer, 2007, 40 (12): 33 - 37.

[8] FAN X, WEBER W D, BARROSO L A. Power provisioning for a warehouse-sized computer [J]. ACM SIGARCH Computer Architecture News, 2007, 35 (2): 13 - 23.

[9] PELLEY S, MEISNER D, WENISCH T F, et al. Understanding and abstracting total data center power [C] //Workshop on Energy-Efficient Design. 2009, 11: 1 - 6.

[10] CHHARIA A, MEHTA N, GUPTA S, et al. Recent trends in artificial intelligence-inspired electronic thermal management [J]. arXiv preprint arXiv: 2112. 14837, 2021.

[11] ISMAIL L, MATERWALA H. Computing server power modeling in a data center: Survey, taxonomy, and performance evaluation [J]. ACM Computing Surveys (CSUR), 2020, 53 (3): 1 - 34.

[12] AKBAR S, LI R, WAQAS M, et al. Server temperature prediction using deep neural networks to assist thermal-aware scheduling [J]. Sustainable Computing: Informatics and Systems, 2022, 36: 100809.

[13] 王继业，周碧玉，张　法，等. 数据中心能耗模型及能效算法综述 [J]. 计算机研究与发展，2019，56 (8)：1587 - 1603.

[14] 余潇潇，马玉草，宋福龙，等. 数据中心能耗建模及能量调节综述 [J]. 电力信息与通信技术，2022，20 (8)：28 - 49.

[15] 罗　亮，吴文峻，张　飞. 面向云计算数据中心的能耗建模方法 [J]. 软件学报，2014，25 (7)：1371 - 1387.

[16] JIN C, BAI X, YANG C, et al. A review of power consumption models of servers in data centers [J]. Appl. Energy, vol. 265, May 2020, Art. no. 114806.

[17] JAYAPRAKASH S, NAGARAJAN M D, PRADO R P, et al. A systematic review of energy management strategies for resource allocation in the cloud: Clustering, optimization and machine learning [J]. Energies, 2021, 14 (17): 5322.

[18] ROY S, RUDRA A, VERMA A. An energy complexity model for algorithms [C] //Proceedings of the 4th conference on Innovations in Theoretical Computer Science. 2013: 283 - 304.

[19] GE R, FENG X, CAMERON K W. Modeling and evaluating energy-performance efficiency of parallel processing on multicore based power aware systems [C] //2009 ieee international symposium on parallel & distributed processing. IEEE, 2009: 1 - 8.

[20] SONG S L, BARKER K, KERBYSON D. Unified performance and power modeling of scientific workloads [C] //Proceedings of the 1st International Workshop on Energy Efficient Supercomputing. 2013: 1 - 8.

[21] LEWIS A W, GHOSH S, TZENG N F. Run-time Energy Consumption Estimation Based on Workload in Server Systems [J]. HotPower, 2008, 8: 17 - 21.

[22] LI Y, WANG Y, YIN B, et al. An online power metering model for cloud environment [C] //2012 ieee 11th international symposium on network computing and applications. IEEE, 2012: 175 - 180.

[23] XU X, TERAMOTO K, MORALES A, et al. Dual: Reliability-aware power management in data centers [C] //2013 13th IEEE/ACM international symposium on cluster, cloud, and grid computing. IEEE, 2013: 530 - 537.

[24] ELNOZAHY E N, KISTLER M, RAJAMONY R. Energy-efficient server clusters [C] //International workshop on power-aware computer systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002: 179 - 197.

[25] TUDOR B M, TEO Y M. On understanding the energy consumption of arm-based multicore servers [C] //Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems. 2013: 267 - 278.

[26] GAO Y, GUAN H, QI Z, et al. Quality of service aware power management for virtualized data centers [J]. Journal of Systems Architecture, 2013, 59 (4 - 5): 245 - 259.

[27] BASMADJIAN R, ALI N, NIEDERMEIER F, et al. A methodology to predict the power consumption of servers in data centres [C] //Proceedings of the 2nd international conference on energy-efficient computing and networking. 2011: 1 - 10.

[28] RIVOIRE S, RANGANATHAN P, KOZYRAKIS C. A comparison of high-level full-system power models [J]. HotPower, 2008, 8 (2): 32 - 39.

[29] BELOGLAZOV A, BUYYA R, LEE Y C, et al. A taxonomy and survey of energy-efficient data centers and cloud computing systems [J]. Advances in computers, 2011, 82: 47 - 111.

[30] FAN X, WEBER W D, BARROSO L A. Power provisioning for a warehouse-sized computer [J]. ACM SIGARCH computer architecture news, 2007, 35 (2): 13 - 23.

[31] TANG C J, DAI M R. Dynamic computing resource adjustment for enhancing energy efficiency of cloud service data centers [C] //2011 IEEE/SICE International Symposium on System Integration (SII). IEEE, 2011: 1159 - 1164.

[32] YAO Y, HUANG L, SHARMA A, et al. Data centers power reduction: A two time scale approach for delay tolerant workloads [C] //2012 proceedings ieee infocom. IEEE, 2012: 1431 - 1439.

[33] LEFURGY C, WANG X, WARE M. Server-level power control [C] //Fourth International Conference on Autonomic Computing (ICAC'07). IEEE, 2007: 4 - 4.

[34] GUPTA V, NATHUJI R, SCHWAN K. An analysis of power reduction in datacenters using heterogeneous chip multiprocessors [J]. ACM SIGMETRICS Performance Evaluation Review, 2011, 39 (3): 87 - 91.

[35] Info-Tech Research Group. Top 10 energy-saving tips for a greener data center [J]. Operate & Optimize Info-Tech Advisor Premium-operate, 2007, 11.

[36] 杨丽娜, 赵 鹏, 王佩哲. 基于 GRU 神经网络的数据中心能耗预测模型研究 [J]. 电力信息与通信技术, 2021, 19 (3): 10 - 18.

[37] SHIN D, KIM J, CHANG N, et al. Energy-optimal dynamic thermal management for green computing [C] //Proceedings of the 2009 International Conference on Computer-Aided Design. 2009: 652 - 657.

[38] BERTRAN R, GONZALEZ M, MARTORELL X, et al. A systematic methodology to generate decomposable and responsive power models for CMPs [J]. IEEE Transactions on Computers, 2012, 62 (7): 1289 - 1302.

[39] CHEN X, XU C, DICK R P. Memory access aware on-line voltage control for performance and energy optimization [C] //2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE, 2010: 365 - 372.

[40] BASMADJIAN R, DE MEER H. Evaluating and modeling power consumption of multi-core processors [C] //Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet. 2012: 1 - 10.

[41] SHI W, WANG S, LUO B. CPT: An energy-efficiency model for multi-core computer systems [C] //Proc. 5th Workshop Energy-Efficient Des. 2013: 1 - 6.

[42] LIM J, LAKSHMINARAYANA N B, KIM H, et al. Power modeling for GPU architectures using McPAT [J]. ACM Transactions on Design Automation of Electronic Systems (TODAES), 2014, 19 (3): 1 - 24.

[43] HONG S, KIM H. An integrated GPU power and performance model [C] //Proceedings of the 37th annual international symposium on Computer architecture. 2010: 280 - 289.

[44] LIN J, ZHENG H, ZHU Z, et al. Thermal modeling and management of DRAM memory systems [C] //Proceedings of the 34th annual international symposium on Computer architecture. 2007: 312 - 322.

[45] LIN W, SHI F, WU W, et al. A taxonomy and survey of power models and power modeling for cloud servers [J]. ACM Computing Surveys (CSUR), 2020, 53 (5): 1 - 41.

[46] VIJAYKRISHNAN N, KANDEMIR M, IRWIN M J, et al. Energy-driven integrated hardware-software optimizations using SimplePower [J]. ACM SIGARCH Computer Architecture News, 2000, 28 (2): 95 - 106.

[47] STOESS J, LANG C, BELLOSA F. Energy management for hypervisor-based virtual machines [C] // 2007 USENIX Annual Technical Conference. CA, USA: USENIX Association Berkeley, 2007: 1 - 14.

[48] ALLALOUF M, ARBITMAN Y, FACTOR M, et al. Storage modeling for power estimation [C] //Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference. 2009: 1 - 10.

[49] VASQUES T L, MOURA P, DE ALMEIDA A. A review on energy efficiency and demand response with focus on small and medium data centers [J]. Energy Efficiency, 2019, 12: 1399 - 1428.

[50] JANACEK S, SCHR DER K, SCHOMAKER G, et al. Modeling and approaching a cost transparent, specific data center power consumption [C] //2012 International Conference on Energy Aware Computing. IEEE, 2012: 1 - 6.

[51] ZHANG X, LU J J, QIN X, et al. A high-level energy consumption model for heterogeneous data centers [J]. Simulation Modelling Practice and Theory, 2013, 39: 41 - 55.

[52] GUAZZONE M, ANGLANO C, CANONICO M. Exploiting VM migration for the automated power and performance management of green cloud computing systems [C] //Energy Efficient Data Centers: First International Workshop, E 2 DC 2012, Madrid, Spain, Mai 8, 2012, Revised Selected Papers 1. Springer Berlin Heidelberg, 2012: 81 - 92.

[53] MCCULLOUGH J C, AGARWAL Y, CHANDRASHEKAR J, et al. Evaluating the effectiveness of model-based power characterization [C] //USENIX Annual Technical Conf. 2011, 20: 19 - 20.

[54] MAKRIS T. Measuring and analyzing energy consumption of the data center [Z]. 2017.

[55] LIANG LUO, W. U. WEN-JUN, AND FEI ZHANG. Energy modeling based on cloud data center, Journal of Software, 2014 (7): 1371 - 1387.

[56] YANG H, ZHAO Q, LUAN Z, et al. iMeter: An integrated VM power model based on performance profiling [J]. Future Generation Computer Systems, 2014, 36: 267 - 286.

[57] HARTON T W, WALKER C, O'SULLIVAN M. Towards power consumption modeling for servers at scale [C] //2015 IEEE/ACM 8th International Conference on Utility and Cloud

Computing (UCC). IEEE, 2015: 315 - 321.
[58] KE G, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree [J]. Advances in neural information processing systems, 2017, 30.
[59] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine [J]. Annals of statistics, 2001: 1189 - 1232.
[60] LIN W, WANG H, ZHANG Y, et al. A cloud server energy consumption measurement system for heterogeneous cloud environments [J]. Information Sciences, 2018, 468: 47 - 62.
[61] LI Y, HU H, WEN Y, et al. Learning-based power prediction for data centre operations via deep neural networks [C] // Proceedings of the 5th International Workshop on Energy Efficient Data Centres. 2016: 1 - 10.
[62] LIU N, LIN X, WANG Y. Data center power management for regulation service using neural network-based power prediction [C] //2017 18th International Symposium on Quality Electronic Design (ISQED). IEEE, 2017: 367 - 372.
[63] LIN W, WU G, WANG X, et al. An artificial neural network approach to power consumption model construction for servers in cloud data centers [J]. IEEE Transactions on Sustainable Computing, 2019, 5 (3): 329 - 340.
[64] ROY N, DUBEY A, GOKHALE A. Efficient autoscaling in the cloud using predictive models for workload forecasting [C] //2011 IEEE 4th International Conference on Cloud Computing. IEEE, 2011: 500 - 507.
[65] CHEN Z, ZHU Y, DI Y, et al. Self-adaptive prediction of cloud resource demands using ensemble model and subtractive-fuzzy clustering based fuzzy neural network [J]. Computational intelligence and neuroscience, 2015, 2015: 17 - 17.
[66] DHIMAN G, MIHIC K, ROSING T. A system for online power prediction in virtualized environments using gaussian mixture models [C] //Proceedings of the 47th Design Automation Conference. 2010: 807 - 812.
[67] ZHU H, DAI H, YANG S, et al. Estimating power consumption of servers using gaussian mixture model [C] //2017 Fifth International Symposium on Computing and Networking (CANDAR). IEEE, 2017: 427 - 433.
[68] MYDHILI S K, PERIYANAYAGI S, BASKAR S, et al. Machine learning based multi scale parallel K-means++ clustering for cloud assisted internet of things [J]. Peer-to-Peer Networking and Applications, 2020, 13: 2023 - 2035.
[69] SHAHIDINEJAD A, GHOBAEI-ARANI M, MASDARI M. Resource provisioning using workload clustering in cloud computing environment: a hybrid approach [J]. Cluster Computing, 2021, 24 (1): 319 - 342.
[70] SALLAB A E L, ABDOU M, PEROT E, et al. Deep reinforcement learning framework for autonomous driving [J]. IS&T Electronic Imaging, Autonomous Vehicles and Machines 2017, AVM-023, 2017: 70 - 76.
[71] TUCKER A, GLEAVE A, RUSSELL S. Inverse reinforcement learning for video games [J]. arXiv preprint arXiv: 1810.10593, 2018.
[72] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: A survey [J]. The International Journal of Robotics Research, 2013, 32 (11): 1238 - 1274.
[73] FARAHNAKIAN F, LILJEBERG P, PLOSILA J. Energy-efficient virtual machines consolidation in cloud data centers using reinforcement learning [C] //2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing. IEEE, 2014: 500 - 507.
[74] LIN X, WANG Y, PEDRAM M. A reinforcement learning-based power management framework for green computing data centers [C] //2016 IEEE International Conference on Clo-Engineering (IC2E). IEEE, 2016: 135 - 138.
[75] LIU N, LI Z, XU J, et al. A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning [C] //2017 IEEE 37th international conference on distributed computing systems (ICDCS). IEEE, 2017: 372 - 382.
[76] SHAKYA S, SHAKYA S. Resource allocation and power management in cloud servers using deep reinforcement learning [M] //Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021. Singapore: Springer Singapore, 2021: 789 - 798.
[77] TARAFDAR A, SARKAR S, DAS R K, et al. Power Modeling for Energy-Efficient Resource Management in a Cloud Data Center [J]. Journal of Grid Computing, 2023, 21 (1): 10.

(上接第 6 页)

[29] 范天祥. 导弹伺服系统虚拟样机仿真与验证 [D]. 哈尔滨：哈尔滨工业大学，2020.
[30] 谭辉桐. 基于 ADAMS 的电动舵机动力学建模与仿真分析 [D]. 武汉：华中科技大学，2016.
[31] 刘　昂，房　凯，郝新锋，等. 数字化环境下雷达装备量产质量控制技术探讨 [J]. 电子质量，2022 (9)：101 - 103.
[32] 孙　宁，周红桥. 军用电子装备三维数字化技术标准体系研究 [J]. 中国标准化，2018 (3)：107 - 113.
[33] 刘国庆，杨雨松，赵宝峰. 基于数字军工的装备质量监督模式创新研究 [J]. 装备学院学报，2014，25 (5)：30 - 33.
[34] 胡长明，梅启元，张　柳，等. 结构工艺样机全流程贯通关键技术研究与应用 [J]. 电子机械工程，2021，37 (3)：1 - 8.
[35] 谭建荣. 数字样机共性关键技术及其应用 [C] //高档数控机床与制造工艺创新论坛论文集，2009：26 - 27.
[36] 罗　旭. 机械产品维修性设计与优化技术研究 [D]. 长沙：国防科学技术大学，2008.
[37] 郭建麟. 电子设备数字样机建模与仿真分析研究 [D]. 天津：天津大学，2008.
[38] 曹　波. 袋式除尘器数字化测试样机集成数据处理及可视化研究 [D]. 天津：河北工业大学，2014.
[39] 陈帝江，张红旗，周红桥，等. 军用电子装备结构数字化设计与制造标准研究 [J]. 标准科学，2013 (4)：48 - 52.
[40] 李士刚，王坤云，袁　烨，等. 复杂装备系统任务可靠性在役考核评估方法 [J]. 空天防御，2023，6 (1)：23 - 28.
[41] 胡小利，白　奕. 武器装备系统数字孪生技术 [J]. 指挥控制与仿真，2023，45 (1)：11 - 14.