DOI 10.15302/J-SSCAE-2025.11.015

人工智能驱动的算力基础设施能效优化技术现状及展望

陈晓红[1,2,3,4]，**郑博文**[1,2,3]，**袁依格**[3*]，**唐鸿凯**[3,4]，**陈汉青**[3,4]

（1. 湖南工商大学前沿交叉学院，长沙 410205；2. 湖南工商大学管理科学与工程学院，长沙 410205；
3. 湘江实验室，长沙 410205；4. 中南大学商学院，长沙 410083；）

摘要：人工智能（AI）推动算力需求持续增长，也使算力基础设施能耗压力与运行成本约束更为突出，以硬件节能、静态策略为主的传统优化方式难以满足复杂运行环境下的能效治理需求。本文围绕AI驱动的算力基础设施能效优化技术，从基础设施层、调度与运行层、跨系统协同层3个维度出发系统梳理了相关研究进展。在基础设施层，总结了能效与碳效评价指标体系、能耗与热行为建模方法、冷却系统智能控制等关键技术，讨论了全生命周期评价在算力场景中的应用拓展。在调度与运行层，归纳了基于负载与能耗预测、强化学习与多目标决策的资源调度及功率管理方法，强调在满足服务质量约束前提下实现能耗、性能、碳排放的综合权衡。在跨系统协同层，综述了碳信号感知调度、跨地域算力迁移、算力与能源系统联动优化的研究与工程实践，指出仿真验证、分阶段上线、对照评估对相关策略落地的重要作用。进一步，从数据与模型可信性系统稳定性与可控性、评测标准与可验证体系、工程经济性与政策机制协同等方面归纳了AI驱动的算力基础设施能效优化技术挑战，展望了从芯片到系统的一体化能效管理、能碳协同调度框架、边缘与分布式算力能效治理、标准化评价体系建设实验仿真与工程实践融合等未来方向，作为算力基础设施能效优化研究和应用的系统性构思与参考。

关键词：人工智能；算力基础设施；能效优化；资源调度；机器学习

中图分类号：TD67；TD82　　**文献标识码：**A

AI-Enabled Energy-Efficiency Optimization for Computing Infrastructure: State of the Art and Future Directions

Chen Xiaohong[1,2,3,4], **Zheng Bowen**[1,2,3], **Yuan Yige**[3*], **Tang Hongkai**[3,4], **Chen Hanqing**[3,4]

(1. School of Advanced Interdisciplinary Studies, Hunan University of Technology and Business, Changsha 410205, China; 2. School of Management Science and Engineering, Hunan University of Technology and Business, Changsha 410205, China; 3. Xiangjiang Laboratory, Changsha 410205, China; 4. Business School, Central South University, Changsha 410083, China)

Abstract: The continued growth of computing demand driven by artificial intelligence (AI) has intensified energy consumption pressures and operating cost constraints in computing infrastructure. Conventional approaches dominated by hardware-level energy saving and static policies are increasingly insufficient for energy-efficiency governance under complex and dynamic operating

收稿日期：2025-11-13；**修回日期：**2025-12-30

通讯作者：*袁依格，湘江实验室副研究员，研究方向为人工智能、数智营销；E-mail: immyyuan23@163.com

资助项目：国家自然科学基金项目(72088101)；中国工程院咨询项目“新一代信息技术赋能的数字生态文明建设战略研究”(2023-JB-09)；湘江实验室项目(24XJ01001, 23XJ01002)

本刊网址：sscae.engineering.org.cn

conditions. This study reviews the research progress in AI-enabled energy-efficiency optimization for computing infrastructure from three perspectives: infrastructure layer, scheduling and operating layer, as well as cross-system coordination layer. At the infrastructure layer, we summarize key techniques including energy- and carbon-efficiency metric systems, power and thermal behavior modeling, and intelligent cooling control, and analyze the extension of life-cycle assessment to computing scenarios. At the scheduling and operating layer, we review resource scheduling and power management methods based on workload and energy prediction, reinforcement learning, and multi-objective decision making, emphasizing integrated trade-offs among energy consumption, performance, and carbon emissions under service-quality constraints. At the cross-system coordination layer, we survey research and engineering practices on carbon-signal-aware scheduling, cross-region workload migration, and coordinated optimization between computing and energy systems, highlighting the importance of simulation-based validation, phased rollout, and controlled comparisons for practical deployment. Furthermore, we summarize major challenges related to data and model trustworthiness, system stability and controllability, evaluation standards and verifiable benchmarking, and techno-economic considerations and policy coordination. Finally, we outline future directions including chip-to-system integrated energy-efficiency management, energy–carbon co-optimization frameworks, energy-efficiency governance for edge and distributed computing, development of standardized evaluation frameworks, and integration between simulation and engineering practices, thereby providing a systematic reference for research and deployment of energy-efficiency optimization in computing infrastructure.

Keywords: artificial intelligence; computing infrastructure; energy efficiency optimization; resource scheduling; machine learning

一、前言

自生成式人工智能（AIGC）应用落地后，云计算、大数据、物联网、数字孪生等支撑技术得到二次发展，全球算力基础设施的规模与复杂度进一步提升[1,2]。算力体系的持续建设推动了数据中心、超级计算平台、“云–边–端”协同系统的稳步发展，但随之而来的是能源消耗激增、资源负荷攀升趋势明显[3]。国际能源署（IEA）预测，全球数据中心用电量将在2030年前增长至945 TW·h/年，较当前水平几乎翻倍[4]。我国数据中心用电规模将保持年均约15%的增长，算力系统能耗压力持续上升[5]。美国数据中心电力负荷将从当前的35 GW增加至2035年的78 GW[6]。能源消耗正在成为限制算力系统规模化发展的结构性瓶颈，而传统上依赖硬件叠加、静态资源配置的算力扩容模式难以获得显著的能效收益，算力增长与能耗线性攀升的发展逻辑逐步被打破，亟待构建新的能效提升范式。在此背景下，人工智能（AI）成为算力基础设施能效演进的重要驱动力。多个国家启动了智能化算力运营体系建设，通过AI负载预测、智能调度、能耗自适应管理等方式提升资源利用率，探索低碳、智能、高效的算力体系新技术路径[7,8]。这一发展趋势反映出算力系统将逐步由算力堆叠向AI算力运营过渡，兼顾性能提升与能源成本约束，支撑数字基础设施可持续演进[9,10]。

算力基础设施的电力需求增长远超传统产业，相关的能源结构与运行方式备受关注。以数据中心为例，电力消耗约占全球年度总用电量的1%，相应的温室气体排放量约占全球总量的2%[11,12]；美国数据中心的平均碳强度为548 gCO_2/(kW·h)，显著高于美国其他行业的平均值[13]。高性能计算（HPC）系统亦有类似情况，运行阶段的碳排放量约占全生命周期的87%[14]。在直接用电之外，数据中心需为制冷、备用电源、配电损耗等支出大量能量，控温系统还可能因结构老化或实时监控不足而带来额外损耗。更为关键的是，算力扩展和代际升级带来的设备更新与功率密集度提升，正在削弱硬件效率提升的边际收益[15,16]。从资源利用的角度看，算力资源调配仍较多依赖静态规划、缺乏负载动态优化策略，致使处理器、存储器、网络设备等难以在峰谷交替中得到充分利用，从而放大了空闲功耗与整体能效低下的问题。这表明，当前算力基础设施存在规模扩张带来能耗快速提升的情况，面临电力供给结构性依赖、系统损耗放大、资源动态管理滞后的多重挑战。亟需探寻算力基础设施能效优化的实践应用方法，而AI算法成为各方研究的关注点[17]。

从应用原理上看，AI介入算力能效优化并非简单地叠加算法，而是围绕感知、预测、优化、控制等主要环节重塑运行机制：在感知层，以遥测、数字孪生为基础，采集功率、温度、负载、电网碳强度等关键变量[18]；在预测层，利用监督学习、深度时序模型刻画负载、功耗、热行为的耦合关系，支持对数据中心的电源使用效率（PUE）、热点分布、峰值功率等的提前预测；在优化层，将能耗与碳排放显式纳入目标函数，应用深度强化学习（DRL）、多目标调度策略，实现性能、能效、服务等级协议（SLA）等的优化[19]；在控制层，将决策下沉至冷却

系统、动态电压与频率调节（DVFS）/ 电源状态管理、容器与作业编排器、碳感知的时空迁移，开展冷却参数、频率电压、负载分布、跨机房任务迁移等的联动控制[20,21]。上述措施在设施端、系统端都产生可量化收益：通过强化学习（RL）与预测控制，显著降低数据中心的的制冷能耗；面向调度与编排的碳感知平台，将电网碳强度预测、作业可推迟性评估、风险约束优化集成为可落地的生产系统[22]；在更细分的层级上，覆盖 PUE / 能耗建模[23]、DVFS 与工作负载感知[24]、跨数据中心碳敏感流量 / 数据迁移[25]等场景，逐步将以算力换性能的“红色”范式推向以智能换效率的“绿色”范式。上述路径与硬件演进、结构改良兼容，通过数据驱动的自适应控制，将分散的节能点优化并耦合到可持续的系统一体化调度，在不牺牲可用性、服务质量的前提下扩大能效与减排的可实现边界。

当前已有针对数据中心能耗预测与调度优化的不少研究，但多聚焦单一层面的能耗预测或资源调度，缺乏对 AI 在算力基础设施能效优化中系统性作用机制的梳理[26]。本文围绕算力基础设施能效优化，从基础设施层、调度与运行层、跨系统协同层 3 个维度出发，梳理 AI 驱动算力基础设施能效优化的关键技术进展；立足预测控制、RL、生命周期评价等方法的应用实践，归纳跨层协同优化、电算协同运行等技术特征，进而探讨 AI 一体化能效管理、能碳协同调度、边缘与分布式能效治理、标准化评测体系建设、实验仿真与工程实践融合等未来发展方向。

二、算力基础设施的能耗特征、能效评估体系与应用阶段

算力基础设施是数字经济、AI 发展的关键支撑，由计算、存储、网络、供配电、冷却等核心设备构成，运行形态呈现高密度集成、高并发负载、长周期稳定运行等特征（见表 1）[27]。算力中心不仅是传统的数据处理设施，而且演化为集 HPC、海量数据存储、智能调度、能源管理于一体的综合系统。图形处理器（GPU）、AI 加速节点带来峰值功耗与热密度的增长，显著推升供配电与散热压力；网络、存储、制冷等支撑环节在算力规模扩大过程中的占比不断上升，使能耗结构呈现多元化与系统化特征。在此背景下，能耗成为决定算力系统可持续能力的关键约束因素，有必要从组成单元出发，对算力基础设施的能耗构成、能效评价体系进行深入分析，以揭示不同组件的能耗特征与能效瓶颈，为后续的节能降碳策略、AI 驱动的自适应能效管理等提供理论依据。

（一）能耗构成分析

算力基础设施的能耗密度高、耦合强、结构分层，来源包括服务器计算节点、网络互联、存储集群、供配电、冷却系统等环节，具有信息技术（IT）设备能耗、基础设施能耗的双重特征[28]。① 在系统层面，GPU/AI 加速节点是能耗的主力形式，单节点功耗为 700~1200 W，高密度训练集群甚至超过 100 kW/机柜，计算能耗占总能耗的 45%~

表 1　算力基础设施的主要设备与作用

设备类别	核心设备 / 关键技术构成	技术作用
计算与加速节点	GPU/AI 专用集成电路服务器	提供大规模并行算力，支持模型训练与推理
通用计算与调度	CPU 服务器	承担任务调度、数据预处理和部分推理工作
高速网络	服务器 NIC、核心 / 汇聚交换机（以太网 / InfiniBand）	节点间高速互联与通信，支持训练集群通信
分布式存储	全闪块 / 文件 / 对象存储系统	存放训练数据、模型检查点，支持高并发访问
本地高速存储	NVMe SSD（节点本地缓存）	提升数据本地访问性能，降低远程读写延迟
系统互连与内存	NVLink / NVSwitch、PCIe Gen5、CXL 内存扩展	提供节点内高带宽互连、可扩展内存资源
供配电	UPS、机架电源分配单元、服务器电源供应单元	实现稳定供电、能耗分配、电力计量
冷却系统	房级风冷、直冷液冷（冷板+CDU）	设备散热与热管理，保障系统稳定运行
机柜与通道管理	机架系统、冷热通道封闭方案	物理承载、气流优化、空间管理
监测与计量	电力分项计量、环境传感器	能耗监测、环境状态采集、运行可视化

注：CPU 表示中央处理器；NIC 表示网络接口卡；NVMe 表示非易失性内存主机控制器接口规范；SSD 表示固态硬盘；NVLink 表示 NVIDIA 公司推出的总线及其通信协议；NVSwitch 表示 NVIDIA 公司推出的互联交换设备；PCIe Gen5 表示第 5 代高速串行计算机扩展总线标准；CXL 表示计算机高速互连；UPS 表示不间断电源；CDU 表示冷却分配单元。

55%[29]。② 为了避免高热通量导致节点性能退化甚至失效，冷却系统提供持续散热能力，相应能耗在HPC、AIGC训练场景中的占比为30%~45%；液冷较风冷可显著降低冷却能耗，但需配置CDU、泵、换热器等辅助设备[30]。③ 计算节点的高功率密度、深度负载波动等，会诱发电力链路冗余和热管理额外开销（如电源冗余设计、UPS转换损耗、服务器电源模块效率下降等），使供配电部分的能耗占总能耗的10%~15%[31]。④ 网络与存储系统能耗随集群规模呈指数级增长，大型AI训练系统采用基于以太网／InfiniBand的高速网络、NVMe分布式存储，相关的交换芯片、NIC、SSD阵列的合计能耗占总能耗的10%~20%且呈抬升趋势[32]。

（二）能效评价指标与模型

从能效评价体系看，PUE（数据中心总耗电量／IT设备耗电量）、碳使用效率（CUE=数据中心年碳排放量／IT设备能耗）是分别衡量算力设施运行效率、碳排放强度的核心指标。PUE越接近1，表示能源利用效率越高，电力主要用于实际计算而非冷却、照明、配电等附属系统；CUE反映消耗1 kW·h电力产生的碳排放量。目前，数据中心采用先进液冷后PUE值可稳定在1.1~1.2，显著优于传统的风冷架构[33]。然而，在高密度GPU集群、实时推理等业务场景中，热负载、通信密度、电力冗余需求会同步提升，表明能耗并非单纯随计算规模以线性方式增长，而是伴随散热需求、供能链路压力、网络通信功耗等呈系统性增长[34]。

数据中心基础设施效率（DCiE）、总碳效率（TCE）等指标与PUE、CUE共同构成数据中心能效评估的核心指标体系。DCiE是PUE的倒数，定义为IT设备能耗占总能耗的比例，用于直观衡量能量利用效率[35]。TCE综合反映电力结构与碳排放强度，将能源消耗、设备效率、碳转化系数纳入统一框架，用于表征单位计算量引发的碳排放水平[36]。此外，随着云计算、边缘计算协同的扩展，有研究提出了考虑能源复用率、水资源利用效率的复合指标，以反映不同的地域、能源类型、制冷方式下的能效差异[37]。

从系统层角度看，生命周期能耗分析是衡量算力设施可持续性能的有效方法，对建设、运营、维护、退役阶段的能耗与碳排放进行全流程核算，识别出能源消耗与环境负担的主要来源[38]。在数据中心全生命周期的碳排放中，约80%~90%位于运行阶段，以冷却、供电系统为最主要的排放环节；基础设施制造与运输环节占比有限，但在高密度GPU计算节点中呈上升趋势[39]。IEA等机构建议在新建和扩容阶段采用设计阶段碳评估方法，将建筑材料碳强度、设备能效等级、电网碳因子等纳入综合评估体系[40]。可见，能效优化不再局限于运行环节，而应在规划与建设阶段即融入全生命周期碳管理理念。

在任务层和算法层，AI计算负载的能耗测度模型逐渐成为研究关注点，如尝试建立从浮点运算次数（FLOPs）到能耗与碳排放强度的推导模型，支持开展计算任务级的能效分析[41,42]。根据谷歌公司的估算，训练拥有1.75×10^{11}个参数的Transformer模型消耗的电能约为1287 MW·h，对应的碳排放超过550 tCO_2[43]。当前，主流方法将能耗E、估算碳排放量C表示为：

$$E=\alpha\times\mathrm{FLOPs}+\beta$$
$$C=E\times\gamma \quad (1)$$

式（1）中，α表示硬件能效系数，β表示设备待机及数据传输功耗，γ表示单位电能的碳排放系数[44]。相关方法为AI任务能耗基准化提供了统一度量框架，也为算力基础设施的能效预测与任务调度优化提供了基本依据。当然，有研究指出此类方法未能区分硬件架构、芯片制程与电源管理等因素带来的差异，还需探索基于RL与自适应控制的实时能耗预测框架以实现更精细的能效调度[43]。

（三）AI在算力设施能效提升中的应用阶段

从发展脉络看，AI在数据中心能效提升中的作用表现为从辅助分析走向闭环控制、从局部节能走向系统协同、从单域优化走向能碳协同治理。早期研究主要依托监测数据，采用回归与统计学习方法对功耗、温度、冷却负载等进行预测及告警，AI更多承担的是能耗认知与运行诊断功能[45]。随着运行数据实时化、负载形态复杂化，深度时序建模与多源融合方法成为主流，推动能耗预测从单变量拟合扩展为针对功耗、热状态、负载的耦合建模，为调度与控制提供先验信息。

在此基础上，RL等方法开始应用于冷却控制与资源调度环节，形成面向安全约束的闭环优化范

式，使 AI 从提供辅助建议转向参与决策及执行[18,24,46]。近年来，在低碳约束强化、算力网络化的背景下，优化目标中进一步纳入电价、可再生能源出力、碳强度信号，形成碳感知调度、跨地域算力迁移等协同策略，使能效优化从能耗最小化拓展为包含性能、能耗、碳排放、成本的多目标权衡[47]。

整体上，AI 在算力设施能效提升中的应用成熟度出现了结构性分化，预测建模较为成熟并获得规模化应用，闭环控制处于从试点走向规模化的过渡阶段，能碳协同仍处于快速演进期（见表 2）。

三、人工智能驱动的算力能效优化技术体系与实践进展

（一）技术体系总体框架

算力基础设施能效优化技术体系是沿着硬件节能驱动、系统级协同优化、AI 驱动闭环治理的发展路径而逐步形成的。算力能效优化路径可概括为从芯片和冷却的底层技术创新扩展到虚拟化与调度机制层面革新，再演进到跨区域协同优化（见图 1）。有必要从技术体系层面出发，对 AI 驱动的算力能

表 2　AI 在算力设施能效提升中的应用情况

阶段划分	时间 / 年	主要目标与应用场景	技术路线	应用进展	成熟度判断
辅助分析与经验建模	2008—2014	监测、告警、容量评估	回归、统计学习、规则与模型融合	基于运行监测数据开展功耗与热状态预测与诊断	工程可用，收益受限
深度预测与多源耦合建模	2014—2019	功耗与热状态预测，支撑调度与冷却优化	LSTM、Transformer、CNN、GNN	多源融合与物理约束增强的耦合建模成为主流方向	预测类规模化成熟
局部闭环控制落地	2016—2021	冷却优化、局部调度	RL、模型预测控制、混合控制	在安全约束下形成可运行的闭环控制范式	试点向规模化过渡
系统级调度与多目标权衡	2019—2024	多租户调度、异构编排	RL、多目标优化、预测与决策耦合	面向动态负载形成自适应调度框架	大规模集群逐步成熟
能碳协同与跨域联动	2021—	碳感知调度、跨区域迁移	碳信号建模、碳感知调度器	纳入电价与碳强度信号的协同调度路径持续涌现	快速演进，待统一框架

注：LSTM 表示长短期记忆网络；CNN 表示卷积神经网络；GNN 表示图神经网络。

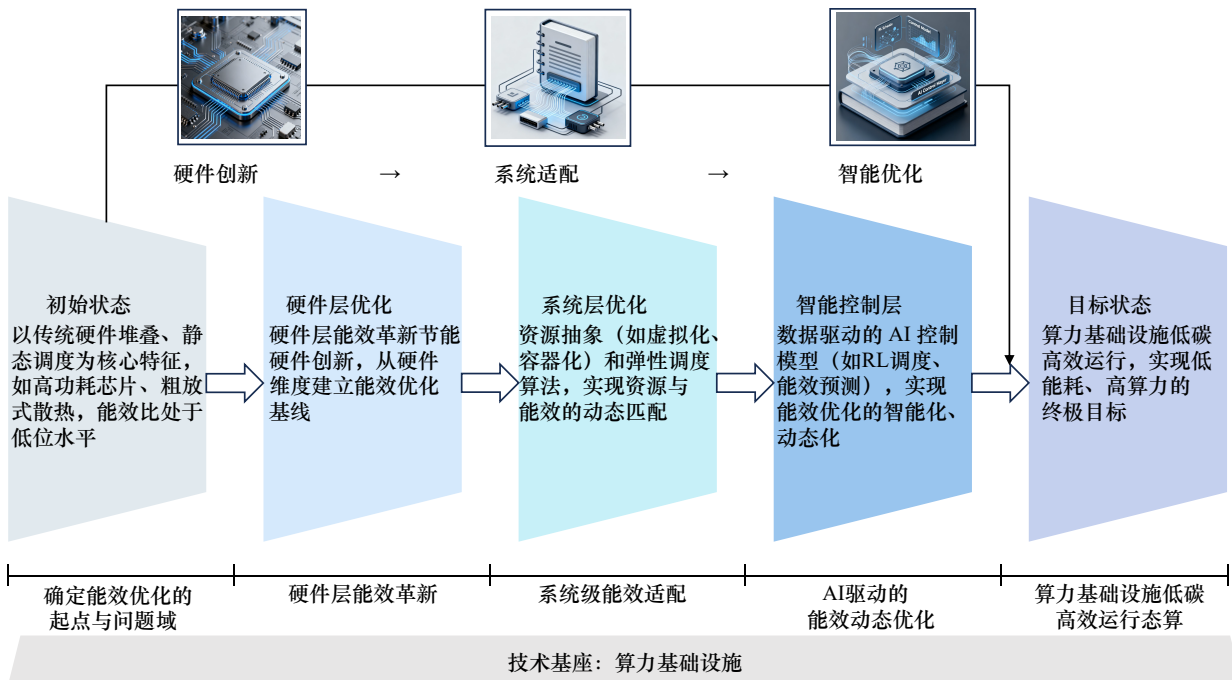

图 1　算力基础设施的能效优化路径

效优化机制进行抽象与整合，以揭示内在逻辑与关键组成。

从演进路径看，早期的算力基础设施能效提升主要依赖硬件层面的节能技术创新，但单纯依靠半导体制程微缩已难以支撑算力需求的持续增长，故业界逐步转向采用异构计算架构、先进冷却技术来实现能耗下降[48]。例如，CPU－GPU异构架构、片上系统设计等在进行任务类型匹配、能效比控制后能够显著降低单位计算能耗，液冷、浸没式冷却、余热回收等技术进一步提升了冷却效率，支持构建了低碳循环算力中心的基本形态[49~51]。随着算力规模扩大、资源利用复杂度提升，系统层优化成为关键路径，如虚拟化、容器化、能耗感知调度等显著提高了集群利用率，通过跨区域任务迁移实现系统级能效优化[52~54]。在此基础上，AI逐步介入算力能耗控制，推动优化模式由规则驱动转向数据驱动、决策驱动[45,46,55,56]。

AI驱动算力基础设施能效优化的整体技术逻辑可抽象为由感知、预测、决策、执行、反馈等构成的闭环结构，具有从数据驱动到决策反馈、从算法智能到架构协同的多层联动特征。在该技术体系下，AI既作为各环节的关键优化工具，又贯穿算力基础设施运行的全生命周期，通过数据感知、模型建构、智能调度、自适应控制，支持从能耗监测到动态优化的系统级能效治理[57~59]。从功能结构的角度，该技术体系可划分为能耗预测与建模、智能调度与资源优化、冷却与环境控制、芯片与算力架构支撑4个相互耦合的层次，各层通过数据流与控制流的双向交互共同构成算力系统的智能能效治理体系。

在感知与预测层，能耗建模是智能调度与系统级优化的基础。LSTM、Transformer等深度时序模型在多节点系统能耗预测中具有显著优势，CNN、GNN等结构用于刻画冷却气流与热流路径，使物理约束与AI算法相结合的能耗、温度、负载耦合模型具有良好的预测精度及结果可解释性[57,59~62]。

在决策与优化层，RL和多目标优化成为AI调度系统的核心方法。基于RL的调度策略可在动态负载条件下实现能耗显著降低并保持服务质量稳定的目标，多目标优化方法在能耗、性能、延迟之间提供了有效的权衡机制[18,24,46,63~67]。将预测与决策进行紧密耦合，使系统具备面向复杂运行环境的实时优化能力。

在执行与反馈层，AI应用于冷却与环境控制，促成算力系统与冷却系统的动态协同。基于DRL与预测控制的冷却优化策略显著降低了PUE和冷却能耗，实现覆盖预测、决策、执行、反馈等过程的智能控制闭环[62,68,69]。从底层支撑看，AI辅助芯片设计、异构算力架构优化、芯粒与光互连等技术，为算力能效治理提供持续演进的硬件基础[70,71]。

（二）能耗预测与建模技术

在算力基础设施的能效管理方面，能耗预测是开展动态调度与智能控制的前提，相关模型如表3所示。传统方法主要有统计回归、经验建模，可以捕捉单变量趋势，但不适用高维度、非线性、时变性强的算力负载场景[58]。随着数据中心运行数据采集的实时化、多源化，深度学习模型凭借良好的时

表3　不同能耗预测模型的比较

模型类型	代表方法	主要输入特征	优点	局限性	典型应用
传统统计模型	ARIMA、SVR、MLR	CPU利用率、负载率	实现简单、可解释性强	难处理高维非线性关系	小型数据中心能耗趋势预测
深度时序模型	LSTM、GRU	CPU/GPU功率、冷却负载	捕捉长时依赖、预测精度高	训练耗时、需大量样本	大型集群能耗预测
注意力机制模型	Transformer，时间融合Transformer	多节点温度、任务队列、负载波动	适应突发负载、泛化性能强	计算资源消耗高	云平台能耗优化
混合深度模型	CNN－LSTM、GNN－LSTM	热分布、机架布局、气流场数据	考虑空间－时间耦合、适配复杂结构	模型复杂度高	超算中心温度与能耗协同预测
物理－AI融合模型	回归+GNN／物理信息神经网络	温度梯度、设备功率、环境参数	精度高、具有可解释性	依赖高质量物理约束	智能冷却系统控制与优化

注：ARIMA表示自回归移动平均模型；SVR表示支持向量回归；GRU表示门控循环单元网络。

序建模与特征表示能力，逐步成为能耗预测的主流工具。在能耗预测与建模方面，引入AI算法，推动能效建模从单一负载驱动转向多模态、多层级协同演进。深度时序模型在刻画长时依赖关系方面具有优势。引入注意力机制的模型在异步负载、突发算力需求场景下泛化能力更强。混合结构模型可融合空间布局和时间演化特征，适用于机房结构、热分布、能耗的耦合建模。此外，物理–AI融合模型将热力学约束、设备功率模型等先验知识嵌入学习过程，可提升预测精度并增强模型的可解释性，是极具潜力的前沿能耗预测方法[57~62]。

数据中心的能耗动态不仅受CPU/GPU利用率的影响，还与温度梯度、气流组织、冷却策略高度耦合[61]。数据驱动的“能耗–温度–负载”耦合建模成为新的研究热点，多源异构数据融合建模也是重要的技术途径，联合运用设备运行日志、环境传感数据、冷却系统参数等精细表征能耗行为，为后续的能效调度与控制提供高置信度输入。例如，基于GNN构建温度传播模型，将机房内部设备抽象为图节点，动态推断热流路径[59]；基于多变量回归并引入物理约束校正，提出功率–热耦合模型，将冷却能耗预测误差控制在5%以内[62]。

当前主流的能耗预测模型以全监督学习范式为主，依赖高质量的标注样本和稳定的运行环境，而跨场景泛化能力和长期自适应能力有限；在复杂运行环境下，传感器分布变化、设备老化、业务负载演化等易引发特征分布偏移，进而削弱模型的预测性能。为此，有研究开始探索多模态建模、自监督学习与RL等前沿方向：应用自监督预训练缓解样本标注不足的问题；通过RL实现模型在动态环境中的在线自适应更新，在预测–调度一体化框架下提升系统级能效[57,59,60]。

（三）智能调度与资源优化机制

在算力基础设施中，任务调度和资源分配是影响系统整体能效与服务质量的核心决策环节。传统调度方法基于启发式规则或线性规划（LP）模型，明确静态优先级或预定义策略来进行任务分配，在负载相对稳定的场景中具有实现简单、可解释性强的优势。然而在多租户、高动态、强不确定性的运行环境下，传统调度方法难以兼顾能耗、性能、响应时延等目标[66]。

随着算力系统复杂度的提升，调度问题从单目标优化演化为多目标决策问题，在不确定环境中对多重优化目标进行动态权衡是关键。RL可通过智能体与环境的持续交互，在未知或部分可观测条件下学习近似最优决策策略，在算力调度中获得较多应用，能够应对负载波动、设备状态变化、能耗预测误差等不确定性因素[18,24,46]。例如，阿里云在张北数据中心部署了DRL调度系统，对服务器功率、任务队列、环境温度进行实时感知，利用策略梯度方法动态调整作业分配，在维持计算吞吐稳定性的前提下降低了综合能耗[72]；百度在线网络技术（北京）有限公司在“飞桨”平台上构建了多智能体RL调度框架，以优化异构节点的资源自治和能量分配的方式改善了系统综合功耗与任务响应速度。

在多目标调度场景中，能效优化需要权衡能耗、性能、时延乃至碳排放。属于早期研究的群智能多目标优化（MOO）方法，如粒子群算法、遗传算法及其混合形式，通过构建帕累托前沿来刻画不同目标偏好下的可行解空间[63,73]。在HPC环境中引入MOO算法，可使系统能耗下降18%~25%，而延迟仅为有限增加[63]。这些方法依赖离线求解或静态权重设定，在快速变化的运行环境中实时性与自适应能力受限。

RL与多目标优化的融合成为智能调度新的研究方向。多目标强化学习（MORL）在RL框架中显式引入了多维回报函数或偏好建模机制，使智能体在动态环境中可同时学习多目标权衡策略，进而弥合算法与目标之间的“割裂”[47]；在处理能耗、性能、时延等相互制约目标时，较传统的群智能算法具有更强的在线决策能力与环境适应性[74]。对比传统启发式方法、群智能多目标优化、RL与MORL等典型范式（见表4）可见，MORL在高动态算力环境中对多重能效目标的在线权衡能力具有综合优势。

碳感知调度则进一步拓展了多目标决策的内涵，将电力碳排放因子、可再生能源出力、任务迁移策略相结合，开展算力供给与能源结构的协同优化。例如，亚马逊云科技公司（AWS）通过跨区域任务迁移来匹配可再生能源供给，微软公司Azure利用RL方法开展冷热节点的分级调度，谷歌云将电网碳排放信号直接纳入算力运行的调度决策过程[75]。智能调度与资源优化机制重在以多目标决策

为导向、RL及其多目标扩展形式为依托，构建应对不确定环境的动态决策框架。AI驱动的调度系统融合负载预测、能耗建模、自适应决策等形成闭环优化结构，为算力系统层面的能效稳定性与低碳运行提供决策支撑。

（四）冷却系统与环境控制智能化

冷却系统是数据中心能耗的第二大来源，相应的智能化水平事关算力基础设施的整体能效。传统的冷却策略依赖经验设定与人工调控，难以应对复杂多变的热负载环境。在AI技术引入后，针对液冷和风冷系统的自适应优化能力发展迅速，以冷却能耗、计算负载动态匹配的方式同步提升了能源利用效率与系统稳定性[69]。例如，实时采集机房的温度、湿度、冷却水流速、服务器功率等维度的数据，应用DRL与预测控制算法进行冷却参数的动态调节，使PUE值显著下降[69]。AI模型能够在动态环境中获得连续优化结果，使冷却系统的作用从被动调节转向主动感知与预测控制（见图2），具体过程为：通过传感器网络进行实时数据的采集与状态感知，由深度学习模型进行热负载预测与冷却需求计算，经由RL控制器动态调整阀门开度、冷却水流速、风机转速，应用反馈通路校正模型偏差并持续优化控制参数[68,69,71]。

未来的冷却与环境控制智能化将沿两条路径并行发展：模型层面的融合优化，将供热通风和空气调节（HVAC）能耗预测模型与AI控制机制深度结合，通过端到端的预测–控制框架进行自

表4　不同智能调度与资源优化方法的比较

方法范式	代表算法 / 框架	决策目标特征	优势	局限	典型场景
传统启发式调度	FCFS、轮询调度、Min–Min算法	单目标或隐式目标（吞吐 / 公平）	实现简单、可解释性强	难以适应动态环境与能效约束	小型集群、负载稳定场景
静态优化模型	LP、MILP、加权遗传算法	单目标或加权多目标	理论完备、解质量可控	依赖先验权重、实时性不足	云平台离线或周期性调度
MOO	粒子群优化、遗传算法–MOO、帕累托前沿	显式多目标权衡	能刻画目标偏好空间	多为离线求解，适应性有限	HPC批处理、多租户规划
RL调度	Q学习、DRL、多智能体RL	单目标或隐式复合目标	适应不确定环境、在线学习	目标权衡需人工设计回报	数据中心动态能效调度
MORL	向量回报RL、偏好条件化RL	显式多目标决策	动态权衡目标、在线自适应	训练复杂、稳定性要求高	高动态算力基础设施
碳感知智能调度	MORL+碳因子 / RL+能源模型	“能耗–性能–碳排放”协同	对接能源系统、低碳导向明确	依赖电力与碳信号质量	超算中心、区域算力网

注：FCFS表示先来先服务调度算法；MILP表示混合整数线性规划。

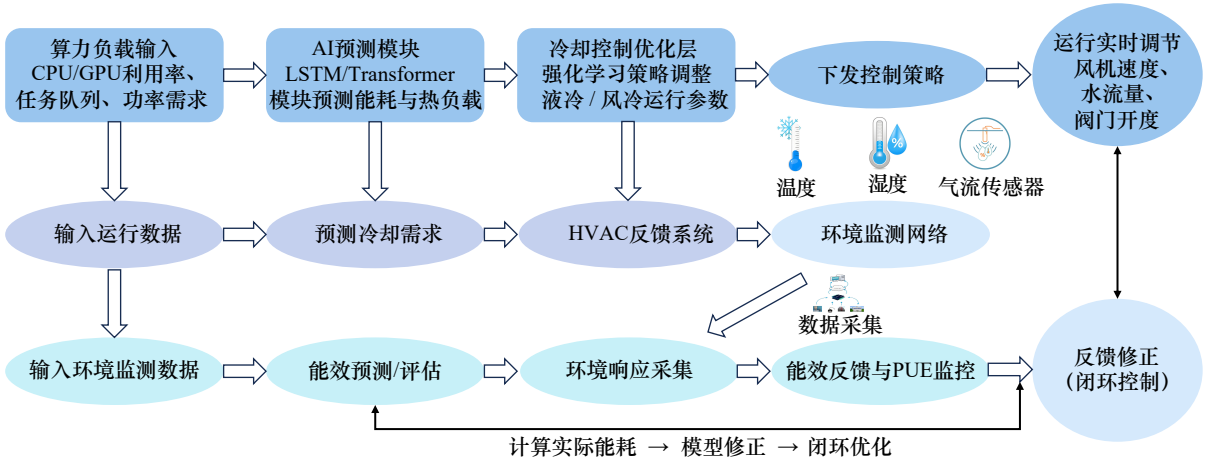

图2　AI驱动的数据中心冷却与环境控制智能化机制

动化的参数调节；系统层面的多源协同，将冷却系统与能耗建模、任务调度等环节联动，使冷却优化转化为算力系统能效闭环过程中的主动环节[68,70,71]。

（五）芯片与算力架构层优化

硬件架构优化仍是进一步突破算力基础设施能效瓶颈的关键。芯片和架构层是支撑AI调度与冷却优化的底层基础，在全栈能效协同体系中发挥着基础性作用。随着摩尔定律逼近极限，算力系统的能效提升逐渐转向架构创新与智能协同[71]。AI在芯片设计、计算架构优化、能效调度方面发挥多重作用，如AI算法支持芯片设计流程优化以降低能耗与功耗密度，芯片架构本身也在针对AI计算任务进行优化（同步推进“AI优化芯片”“为AI设计芯片”）。

在芯片设计环节，AI辅助设计成为提升能效的关键手段，如谷歌公司在芯片布局布线中引入DRL算法，将电路延迟、布线长度分别降低15%、20%，设计周期缩短90%以上[71]。国内企业应用RL、GNN算法来优化电容负载分布，使芯片功耗降低3%~5%；在自动化布线阶段引入多目标优化策略，更好平衡功耗、面积、延迟[70]；自主研发芯片较多采用AI辅助参数搜索与架构设计，显著提升了能效比（部分场景下达到传统通用GPU的1.5倍）[72]。

在计算架构层面，AI驱动的能效优化主要体现在3个方面：发展近存计算架构，将计算单元与存储单元融合，减少数据搬运开销，提升能效2~3倍[71]；采用芯粒模块化设计，将异构算力模块以低功耗互连方式集成，支持灵活调度与按需功耗管理[70]；应用光电混合互连技术，在高密度算力集群中降低通信能耗，为AI调度提供更高的带宽能力[72]。

在系统级协同层面，AI驱动的硬件优化与上层调度、冷却、能耗预测等模块联动，支持从芯片到系统的全栈能效协同。例如，NVIDIA公司通过统一内存访问模型减少CPU和GPU之间的数据传输延迟；华为技术有限公司在昇腾AI架构中实现算力调度层与硬件控制层的能耗协同；阿里云“飞天智算平台”在数据中心引入AI算法，支持跨层功率分配与任务映射，使系统总能耗降低约15%[72,76]。芯片、算力架构层的优化将与AI调度系统深度融合，形成自感知、自学习、自调控的算力基础设施；结合AI辅助设计、异构算力编排、能碳一体化调度，推动从单节点节能到全网能效治理的演进，支持构建安全高效、低碳可控的智能算力底座[70,72,76]。

（六）工程应用与案例

部分互联网企业率先将AI应用到算力基础设施的效能优化，逐步构建了以“感知–预测–决策–反馈”为核心的闭环体系。AI能效优化的优势在于模型驱动与系统集成，通过跨层数据融合与动态调控机制使算力基础设施从被动节能模式演进为主动能效管理。

在数据中心的能耗结构中，冷却系统长期占据较高的比重，运行状态也与IT负载、环境条件强耦合。谷歌公司将DeepMind团队开发的DRL方法引入数据中心冷却系统的控制环节，构建了面向实际运行环境的智能冷却优化框架，实现AI驱动的数据中心冷却系统闭环优化。结合公开信息来看，通过机房内部署的传感器网络，持续采集温度、湿度、气流状态以及IT负载等运行数据，构建了冷却系统运行状态的数字化表征；DRL模型以历史运行数据、实时状态为输入，与仿真环境、真实系统交互，学习不同控制动作（如冷水泵转速、风机功率设定）对能耗、安全约束的影响，形成可在线更新的控制策略。AI控制模块并非完全取代传统控制系统，而是与既有安全机制并行工作，在满足设备安全、温度上限等约束条件下对冷却参数进行优化调节。在长期运行中可使冷却能耗降低约40%，也将数据中心PUE从1.22下降至1.12。该案例验证了DRL模型在复杂工业系统中长期稳定运行的可行性，为AI直接参与基础设施控制提供了工程层面的示范。此外，国内互联网企业利用AI节能算法实现算力与能耗的协同平衡，提出了硬件架构与AI模型深度结合的技术路径。

关于智能算力调度，国内的超算中心保持了由规则驱动向智能自适应体系迈进的技术发展趋势，超算体系逐步向自治计算集群过渡，与国际上AI增强HPC资源编排的主流方向趋同[76]。天津、广州、无锡等超算中心积极推进智能调度工程，将大规模并行任务与异构算力场景作为“试验场”。例

如，在“神威”架构中引入作业画像与任务队列分析模型，通过预测任务计算强度与通信开销来引导节点分配、优化并行划分策略；在“天河”系统上引入资源热度评价指标，将机器学习模型嵌入调度内核，缓解资源热点与排队拥塞；在“神威·太湖之光”系统上开展基于应用行为识别的多粒度资源弹性调控实验，针对万核级作业实现算力、网络、存储方面的自适应协同调度。

国际云服务平台受能耗、低碳算力目标的驱动，不断强化调度智能化策略。以跨地域算力迁移、能源场景感知、AI驱动的调度策略为核心，在为未来算力网络迈向碳感知、能耗预测、自适应资源编排模式确立了工程基础[75]。AWS在EC2集群中采用工作负载特征识别及能耗预测模型，依据能源成本、区域温度、可再生能源供给水平等进行跨区域的实例迁移。Azure构建能耗敏感调度机制，在虚拟化与容器管理层中引入RL策略，支持冷热节点分级调度与弹性扩/缩容。谷歌云依托碳感知调度系统，将不同地区的电网碳排放信号与批处理任务调度相结合，顺应可再生能源的出力节律来调整算力资源供给，协同优化性能、碳排放和成本。

绿色AI算力平台与标准化取得积极进展，国内的工程模式正在引导算力产业从能耗红利转向能效主导的技术逻辑，与国际标准体系发展趋同。我国实施“东数西算”工程，标志着绿色算力体系建设进入工程落地与技术标准体系构建阶段，依据“算随能走、数随绿迁”原则将算力调度与可再生能源基地耦合，推动绿色算力中心集群建设[77]。《2025年度国家绿色数据中心评价指标体系》中包含了能效、碳排放、可再生能源占比指标，形成了涵盖电能计量、冷却效率、虚拟化资源利用率的标准框架[78]。地方性的绿色算力中心示范工程采用区域碳排放系数、能源消纳预测、算网供需动态编排模型，将算力资源向风光资源占比更高区域迁移运行，以液冷系统、高能效服务器、AI能效调控算法共同实现平台级节能。

碳感知算力调度技术发展成为绿色AI基础设施的关键支撑，涉及区域实时碳排放信号采集、计算负载预测、可再生能源供给预测、任务迁移与调度决策优化、执行层的资源治理控制链路等[79]。在碳信号建模方面，构建电力系统碳排放强度时序模型，与数据中心负载模型融合，指导任务错峰与异地调度[80]。在调度优化方面，形成了以碳排放强度预测、任务可延迟性分析、碳约束调度决策为核心的统一框架，如基于RL与多目标优化的跨区域算力迁移方法可在不降级服务质量的前提下取得显著的减排效果[81]。将电力系统“源网荷储”协调机制引入算力系统领域，构建算力、能源双侧需求曲线以及实时碳价格信号，实现算力调度与可再生能源利用的协同优化[82]。通过仿真平台与小规模集群验证，确认了多模态认知计算在碳因子驱动的作业迁移、功率封顶、能效自适应控制策略方面的应用潜力[83]。碳感知调度也从理论模型研究转向实验平台落地阶段，为AI算力系统与低碳能源系统的深度融合提供了方法支撑。

四、人工智能驱动算力基础设施能效优化面临的挑战

（一）存在数据和模型的可信性与可解释性瓶颈

算力基础设施能效优化的核心难题之一是数据质量与语义一致性不足[84]。不同于传统算法，AI能耗优化模型依赖CPU/GPU利用率、功率采样、微环境温度、风冷/液冷动态参数、网络带宽、任务队列日志、区域电网碳强度等运维数据，而实际上数据中心的遥测频率、精度、观测点布局并不统一，数据异常、缺测、噪声等不可避免。不同业务域之间存在访问隔离、隐私保护要求，导致标注稀缺与训练语料不均衡，削弱了模型外推能力，容易产生实验室场景高得分、生产环境低得分的现象[14]。高热密度场景中热传导与能耗的耦合关系具有显著的非线性，仅依靠历史统计难以支撑高置信度预测[22]。

系统稳态优先于理论上的最优态仍是行业共识，运维团队对模型透明度、可信度、解释能力等提出更高的需求，模型的可解释性不足成为工程部署的“心理门槛”与约束[85]。尽管RL与深度网络在负载预测、冷却调控、算力调度中表现优越，但策略产生并非天然可解释，导致运维团队难以掌握关键特征来源和动作触发逻辑。在高可靠计算场景中，黑箱性进一步增加潜在的决策风险，而策略错判可能导致算力调度紊乱、资源争抢与能耗振荡，影响SLA与能源预算执行。

为了应对上述挑战，产业界开始构建可解释能

效模型、联邦算力运维架构，如引入因果推断与机制建模提升决策透明性，通过自监督学习缓解标签不足，采用联邦学习框架和隐私计算实施跨集群协同优化，构建AI辅助与人机共管模式以避免全自动风险[86]。此外，面向IT运营的AI体系逐步内嵌异常检测、模型漂移监控、在线策略回滚机制，确保模型演进可控。整体上，数据治理体系、可解释设计、模型可信工程化成为绿色算力智能化的基础支撑。

（二）面临算法复杂性与系统稳定性矛盾

算力能效优化技术从理论研究走向工程系统的过程中，遭遇算法复杂性、系统稳定性方面的瓶颈。能效调度模型涉及长时序负载预测、多目标优化与实时控制，需在能耗、性能、延迟、可靠性之间取得平衡[87]。尤其是大型GPU集群的任务模式高度动态，算力利用随模型训练阶段、通信与计算占比、网络拓扑的调整而不断变化，使相应待优化问题具有高维非线性特征；高精度模型（如基于Transformer的动态功耗预测）虽然具有前瞻性，但会带来额外的算力开销与调度延迟，可能反向侵占节能收益。在超算场景中，通信拓扑、输入/输出瓶颈、网络拥塞进一步加大了优化复杂性，使单纯追求算法最优易致执行波动与能耗抖动[88]。

调度算法需要在毫秒级响应、分钟级计划、小时级能耗策略之间切换，研发团队面临理论最优、工程可控之间的抉择。企业级平台普遍采用“主控确定性+辅控智能化”架构，即关键调度路径沿用确定性启发式策略（如静态分区、保守功率上限控制），非关键路径部署AI策略进行自适应微调[89]。例如，清华大学与阿里云研究团队在大规模GPU集群上验证了AI辅助功率封顶机制，将模型调参与稳态调度解耦，在关键路径保持确定性策略的前提下，通过轻量化RL策略进行动态功率上限调整，由此平衡能耗与训练收敛速度[90]。鹏城实验室、江苏省算网工程团队在异构算力调度测试平台上引入分阶段灰度验证与安全阈值回退机制，对冷却策略、作业功率限额进行在线微调，确保算力调度在高负载波动场景下维持稳定的服务质量与资源隔离[91]。

实践表明，“主控确定性+辅控智能化”架构可在生产环境中维持高水平的可靠性与服务稳定性，但受限于模型收敛周期、系统安全阈值、灰度上线机制，AI策略的节能潜力通常逐步释放而非一次性发挥。为此，产业界开始探索混合调控与多层时域优化架构，即在顶层引入任务画像与能耗弹性曲线，在底层利用轻量化模型或策略蒸馏进行快速推理，通过图优化与预测缓存来降低调度开销[92]。AI内核开始嵌入稳定性约束与恢复策略（如在线漂移检测、调度回滚、冷却安全边界、负载应急分散机制），以保持专家规则与机器学习的协同。

（三）缺乏标准化与评估体系

算力基础设施绿色化正在从节能工程走向智能运营，但相应的标准与评估体系仍处于早期发展阶段。现行指标体系多聚焦机房物理层面的PUE、CUE等设施向指标，难以精确刻画AI驱动优化产生的系统效益（如任务能效、算力碳强度、调度策略的动态协同收益[93]）。不同厂商的能耗基线、排放因子、功率采样频率等存在差异，使横向比较、跨平台迁移面临障碍（如同一训练任务在不同云平台上的碳排放估算误差为30%~45%[34]）。

在AI能效评价层面，现有的指标体系存在算法优化收益难以独立量化的问题。鉴于AI调度策略涉及功率分配、任务重排、冷却调参、容错补偿等方面的协同动作，相应的节能收益、性能代价难以由单一指标进行全面刻画。产业界开始构建全栈评估框架，纳入算力碳效率、单任务能耗成本、跨时域综合效能等新的指标[94]。还在探索引入生命周期评估、“能耗–碳排放–性能”综合指标体系，将基础设施、调度策略、能源结构纳入统一度量范式[39]。

分别从企业研发、机构监管方面出发，推进标准共建与可验证实验体系建设。数据中心与云平台的碳感知调度基线、能耗数据接口、排放因子标准等正在成形，通过可信执行机制、多方验证环境提升透明度[95]。构建数字孪生算力平台，在虚拟环境中预先模拟能耗行为、碳排放曲线、调度策略等，再采用灰度上线方式实施工程化部署，以降低评估误差并缓解稳定性风险[96]。未来，绿色算力体系将从指标碎片化阶段迈向统一模型、实时指标、动态问责的闭环体系，为智能化调控、系统性进化确立基础条件。

（四）受到工程部署与经济性约束

算力基础设施绿色化从概念验证转向落地应用，面临显著的工程与经济压力。液冷系统、AI调度代理、能耗追踪平台等虽然表现出良好潜力，但设备升级、机房改造、模型运维等加大了资本和运营支出负担[97]。AI驱动的调控系统依赖持续在线训练、实时指标采集、冗余安全策略，导致部署和运维复杂度显著高于静态控制模型[55]。企业预算倾向优先满足算力供给与业务收益目标，而节能策略在资源紧张周期中让位于性能调度，出现绿色策略适配窗口约束。

算力弹性调度、低碳运行机制也面临投资回报周期与即时算力需求的结构性约束。HPC中心、AI训练集群通常采用高功率密度的供电与液冷加速架构，但调度模型（如RL、基于预测的控制策略）引入额外的推理开销与决策不确定性，可能影响延迟敏感任务与大规模训练作业的吞吐表现[98]。产业界倾向采用分层部署架构，由基础控制层提供确定性资源保障与安全阈值，在策略优化层引入可回滚的AI调控模块并通过灰度发布、人机协同运行策略降低上线风险。绿色算力建设也转向渐进式路径，即优先完成软件和调度策略升级再逐步实施冷却系统改造与算力异构化部署，以降低一次性改造成本并缩短投资回收周期。

新兴实践正在引入数字孪生算力系统、投资回报预测模型、碳经济激励机制，以提高部署效率与经济可行性。数字孪生平台可模拟任务调度、热环境响应、能耗行为，为节能策略上线提供可控的验证环境[99]。碳市场价格信号、绿色电力证书制度成为算力投资模型的重要参数，部分算力运营商开始通过区域性绿电优先调度与碳减排收益交易来降低单位算力碳排放并获得增量收益。后续，在政策激励、AI调控成熟度持续提升的基础上，绿色算力将逐步由单纯提升项目效率转型为资本、能源、数字基础设施协同演化的长期投资战略。

五、面向算力系统全生命周期的仿真验证、协同调控与可持续运营框架

随着AI在算力基础设施能效优化中的深入应用，相关研究范式逐步从单点算法突破转向面向真实系统的全链路验证、跨层协同、可持续运行机制建设。构建涵盖仿真验证环境，异构系统协同架构，能源、算力、碳排放一体化的运行机制，成为绿色智能算力发展的关键环节。

（一）复杂算力系统的仿真与验证体系

算力基础设施规模与异构性持续提升，部署能效优化策略已容许黑箱试错[83,86]。不同于互联网业务场景，数据中心和HPC平台对SLA、系统稳定性要求严格，调度或控制策略失误可能引发计算节点宕机、冷却系统失衡、碳成本异常[29,32]。在AI驱动的算力调控体系中，构建覆盖仿真建模、隔离测试、渐进上线的验证流程是能效优化策略工程化落地的前提（见图3）。

当前较为成熟的路径是以数字孪生为核心的仿真验证体系。通过软硬件协同仿真，将服务器微架构、网络链路、冷却系统、电力供给集成至统一的虚拟环境，用于训练与评估能效调控策略[100~102]；提前识别算力负载与能耗的非线性耦合关系，在虚拟环境中测试冷却失效、电力波动、高密度散热异常等高风险工况，降低直接在生产系统中验证策略面临的运行风险。RL与物理模型结合，可以提升复杂工况下环境建模与策略执行的稳定性[18,77]。

对仿真模型与真实系统之间的差异性进行建模是AI调控策略落地应用的关键环节。仿真环境难以完整复现真实系统中的噪声特征、任务波动、传感器误差，导致策略迁移时易出现性能漂移[99]。引入不确定性建模、对抗扰动训练、多场景分布仿真等应对措施，能够缩小仿真与现实系统的差距并提升泛化能力。实践中多采用高密度机架热分布预测、冷却气流扰动模拟、液冷回路延迟建模等，利于AI策略适应机房气流变化、设备热漂移等难以直接观测的动态过程。

在工程部署时，A/B实验（分流测试或对照实验）、灰度发布机制成为算力能效策略上线的通行做法[83]。大型云平台通常采用沙箱测试、小规模集群验证、区域算力池上线、全局调度扩展相结合的分阶段部署方式，与实时回滚与策略审计机制配合，在保障系统可靠性的前提下逐步引入AI调控能力，使能效优化策略稳定融入算力系统的日常运行。

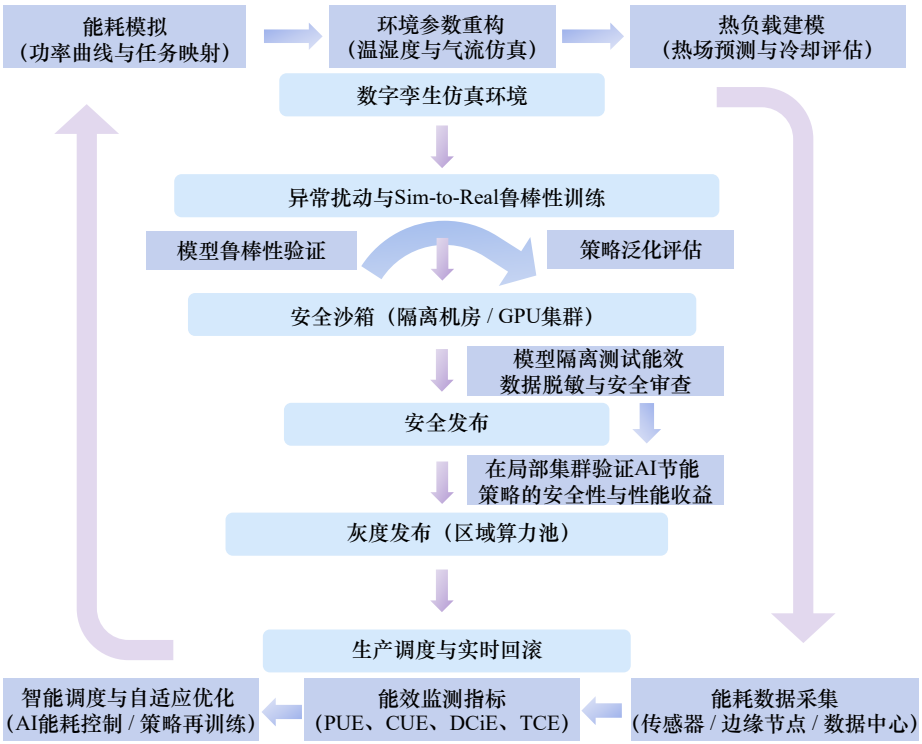

图3 AI 能效策略的上线路径

（二）多层级算力协同与异构系统优化架构

现代算力系统由单一服务器或机房形态演进为覆盖芯片、节点、集群、数据中心、区域算力网络的多层级协同体系，算力消耗结构、系统能效表现均具有显著的层次性，单点智能化优化难以获得系统级能效收益[97]。AI 驱动的算力能效优化需要由局部调节转向跨层协同调控，以在多租户、异构硬件、动态工作负载环境中提升整体能效[15]。

在硬件层面，芯片内部能耗调控趋向与智能策略融合。动态电压与频率调节、芯粒互联控制、片上热通道预测、算力分配策略共同构成微架构能效优化的基础[103,104]。在大规模 AI 推理与训练任务中，利用机器学习预测算子执行特征与内存访问行为，提前优化算子映射与数据通路，从而降低能耗峰值与执行抖动。相关厂商探索在芯片内部引入 AI 代理以实时调节性能与功耗曲线，使芯片由固定功耗策略转向自适应调控[90]。

在节点级与集群级层面，AI 策略逐步嵌入容器编排与异构资源调度体系。容器调度系统、GPU 管理框架、分布式计算平台开始引入时序预测、RL 与多目标规划方法，支持节点算力分配、CPU/GPU 协同调度、存储与网络资源优化[105]。基于作业历史、实时负载特征的调度策略可动态调整任务执行窗口，提升 GPU 利用率并改善集群能效表现[25]。

在数据中心层面，算力负载与冷却系统之间的耦合复杂性成为能效优化的瓶颈之一。AI 可用于风冷路径优化、液冷流量控制、冷冻水系统调度，支持热流分布预测、冷却策略协同优化[22,31]。对任务调度、热管理进行联合建模，以调整算力任务空间分布的方式缓解局部热点、降低冷却系统能耗[102,106]。

在跨区域算力网络层面，算力调度逐步引入能源与碳排放约束。在“东数西算”等工程实践中，建立算力、能源、碳排放耦合模型并与 AI 调度相结合，使负载优先在电价较低、可再生能源占比较高的区域执行，形成跨区域算力分配与能效优化的协同机制（见图4）。类似思路在部分国际云平台中得到应用，使算力调度能够响应区域能源供给条件的变化[19]。

（三）加速绿色算力的经济与政策机制

AI 驱动的能效优化从技术验证转向规模化应用，相应的经济激励与政策体系更显重要。算力基础设施具有初始投入高、生命周期长、能源依赖强

等特征，能效优化涉及技术可行性问题，也事关工程经济性与政策环境适配性[15]。在绿色算力建设过程中，单纯依靠技术性能提升难以持久，需要将能效优化与投资决策、运行成本、收益回收机制相结合，形成覆盖资本支出、运营支出、长期回报的综合评估框架[1]。

在成本收益层面，国际云服务商与HPC中心采用总持有成本、投资回报率（ROI）协同分析方法，开展AI能效技术的经济效果量化评估[75]，分析结果具有先投入、后回收的两阶段特征（见图5）。在部署初期处于资本性支出（CAPEX）主导阶段，液冷改造、传感与数据采集体系、调度平台与模型训练等投入集中释放；随着技术成熟度、部署规模的提升，能耗优化逐步转化为运营费用（OPEX）下降，体现在电费与碳成本下降、峰值电价敏感性降低、运行效率提升等方面[22,23,31,40]。引入能效回报模型，将电费节省、碳配额收益、设备折旧纳入统一核算体系，使节能潜力转化为可量化的经济收益，为不同技术路线优选、投资节奏安排等提供依据[107]。

在政策层面，欧盟碳排放交易体系、美国《通胀削减法案》、中国国家核证自愿减排机制与绿色电力证书制度等，对算力基础设施的能源使用与碳排放形成了外部约束与激励条件。通过碳定价、电力来源认证等方式，将能效表现纳入算力运营的成本与收益结构中，为AI能效优化技术的部署提供了制度环境支撑[58,100,102]。从机制上看，碳市场收益与绿色金融工具的引入均抬升了节能收益曲线或降低了有效投资门槛，使ROI交叉点前移并扩大净收益区间，增强了运营方对液冷改造、能耗预测、智能调度等方案的采用意愿。在工程实践中，经济与政策机制的协同有助于推动算力运营方在满足服务质量要求的基础上，更加系统地评估和采用节能调度、冷热协同、能耗优化等方案，促进绿色算力从技术可行转向经济可行与可持续运行。

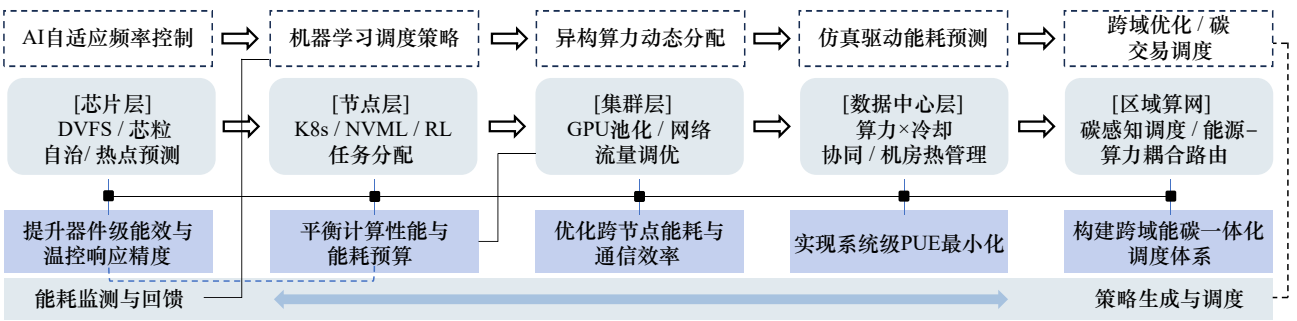

图4　多层级算力协同架构

注：NVML表示NVIDIA管理库。

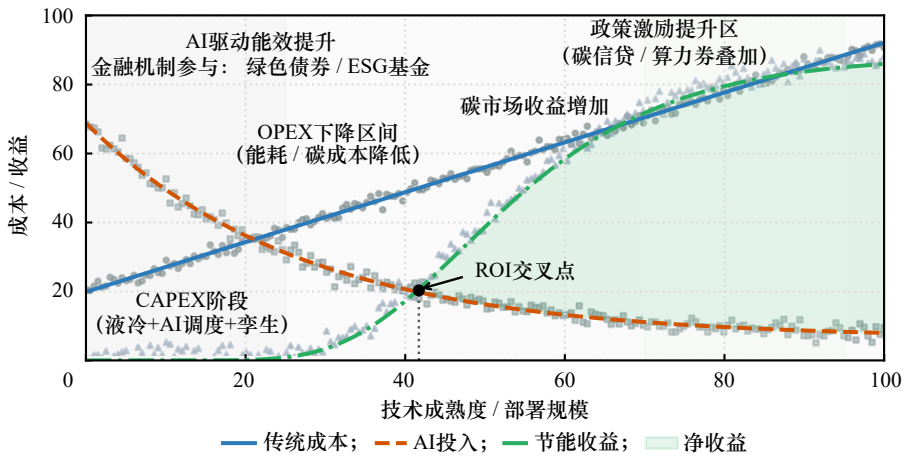

图5　算力节能收益模型

注：ESG表示环境、社会和公司治理。

（四）安全与可信运行保障框架

算力基础设施运行规模扩大、能源耦合加深，算力调度系统在安全属性上接近电力系统、工业控制系统等[15,98]；运行逻辑不能依赖黑箱试错[83,86]，需要构建多层次的安全保障体系，确保调度策略在高复杂、高实时、高弹性环境中具备可追踪性、可回滚能力和故障自愈能力。在AI深度介入调度决策后，系统安全机制由传统的被动告警与人工干预转向融合可验证学习、安全回退、人机协同的运行模式。

在具体实现上，可信AI调度框架成为算力系统安全运行的重要基础。现有研究关注调度策略的可解释性、鲁棒性验证、安全回退机制，以确保在预测偏差、数据漂移、异常负载冲击等情况下系统能够及时切换至规则调度或静态安全模式[91,92]；在数据中心调度体系中引入AI调度沙箱、灰度发布、强制回滚机制[10]，通过策略试运行窗口降低智能调控对生产系统的冲击风险。自监督异常检测、多路径容灾机制逐步成为高可用算力平台的基础配置，用于提升系统在异常工况下的运行韧性。

在算力系统安全运行责任划分层面，逐步引入人机协同、自治等级划分等理念，明确运维人员、AI控制器在不同运行状态下的职责边界。参照自动驾驶分级思路，在较低的自治等级下AI提供决策建议并实施局部能耗优化，在更高的自治等级下AI执行较高比例的资源分配与能耗调控任务并自动处置部分类型的异常状态[107]；在高负载、极端气候、电力异常等关键场景中，运维人员保留最终决策权。由此，算力基础设施安全并非追求完全去人工化，而是在保障系统可控性的前提下，通过人机协同提升调度系统的可靠性与稳定性。

六、算力基础设施能效优化技术研究展望

（一）从芯片到系统的AI一体化能效管理

算力能效优化亟需突破单一层级、单一对象的分析范式，转向覆盖芯片、节点、集群、数据中心、区域算力网络的多层协同建模与联合优化问题研究。目前多在单一层级内开展能效调控，缺乏对不同层级之间的能量流、任务流、控制时序耦合机理的统一刻画[20,25,27]。在芯片层面，开展基于机器学习的负载感知调控机制、传统功耗管理策略的协同设计，尚缺系统性的理论支撑[24,34]。在节点与集群层面，异构算力资源的普及使任务特征、能耗行为与调度决策的映射关系趋于复杂，需要更加通用的跨架构建模方法[41]。在数据中心、区域算力网络层面，尚未形成有效的统一优化框架[2,3]以解决算力调度与冷却系统、能源供给结构、碳排放约束的协同问题。后续，构建跨层耦合的能效建模方法并在不同时间尺度上开展协调控制，是AI驱动算力能效管理的重要研究方向之一。

（二）碳感知与能碳协同调度框架

算力基础设施向低碳运行演化，不再单纯依赖硬件节能与能源替代，而是在运行控制层引入碳感知模型，使算力调度逻辑与能源供给结构、电力运行状态、碳排放约束形成紧密的耦合关系[87]。在全局视角下，算力调度系统逐步纳入电力市场实时价格、区域可再生能源占比、碳排放因子与环境负荷边界，相应的任务分配转向对性能、能耗、碳排放等目标的协调优化[92,94,97]。相应地，算力系统不仅是被动适配能源条件的负载主体，而且开始具备根据电力系统运行信号调整调度策略的能力（如利用实时碳信号将非紧急算力任务迁移至低碳区域或可再生能源充足时段，在数据中心调度中采用算力运行与电力供给协同优化的决策机制[16,17,95]）。在能碳耦合建模层面，将算力任务特征、硬件功耗模型、PUE动态曲线、区域碳强度函数纳入统一优化框架并构建碳感知调度器[23]。在更广泛的能源系统研究中，算力负载被视为具有可迁移性、时间弹性的灵活负荷，相关调度行为关联算力系统自身的能效表现、电力系统的平衡与调节过程。探索将算力基础设施纳入需求响应、虚拟电厂、绿色电力交易机制，使算力调度与电力调度在运行层面协同联动并与碳市场机制配合，形成算力可转移、碳排放可计量、电力可协调的运行体系[80,98]。在此框架下，能碳协同调度不再局限于算力侧的优化，而是呈现向电力系统与算力系统协同运行演进的趋势，成为高比例新能源接入条件下算力发展与能源安全协同的研究课题。

（三）边缘智能与分布式算力能效治理

算力体系从集中式数据中心向大规模分布式边缘节点延伸后，能效治理对象转变为由网关、微型

服务器、移动通信基站、终端计算单元构成的算力网络[2]。依赖集中监控、全局调度的传统优化方法，在节点高度异构、资源受限、运行环境复杂的边缘场景中面临适用性不足的问题。边缘节点面临功耗弹性、散热条件、算力规模方面的约束，相应能效优化更多依赖轻量化模型与本地自治控制机制。现有研究集中在基于时序预测的负载与温度估计、结合轻量RL或分布式模型预测控制的近实时节能策略，但在跨节点协同、模型泛化能力增强、通信开销控制等方面有所不足[27]。任务卸载决策需权衡能耗、通信延迟、服务质量等目标，开展建模的复杂度随节点规模扩展而显著上升[80]。在区域电力碳因子、异构设备能效差异、负载迁移策略相结合，面向低碳目标提出分布式算力调度路径方面已有进展[105]，但多能互补供给、不确定能源状态、跨区域协同控制方面缺乏统一建模与优化框架。在保障边缘服务实时性及可靠性的前提下开展分布式算力与能源系统的协同能效治理，仍是有待深化的关键研究课题。

（四）标准化与评价体系建设

算力能效优化从算法研究走向工程应用后，对指标体系、标准化框架的需求凸显，应发展贯通算力、能耗、碳排放全链条的评价体系[78]。当前，国际上多元标准并存，如绿色数据中心侧强调PUE、CUE等设施级指标[31,40]，云平台引入SLA、执行效率、资源碎片率等运行指标[34]，也有研究提出每任务能耗、每模型训练能耗、碳延迟乘积等算法级指标[97,99]。此外，模型生命周期评价方法向AI算力场景拓展，覆盖训练、推理、模型更新、硬件更新等全生命周期环节，也与碳因子、冷热管理、电力调度耦合，使能效评价从基础设施能耗延伸至算力活动生命周期[58]。然而，不同指标体系在计量边界、数据获取、适用场景上存在差异，导致跨平台比较、绿色算力认证、算力交易仍面临制度与技术障碍；尽管欧盟可持续数据中心认证体系、国内“东数西算”工程已在政策层面推动多维评价框架建设[11,77]，但在异构算力、边缘节点、AI调度模块的统一监测与可验证报告方面仍有明显不足[17,20,23]。后续，构建开放、可验证、可扩展的算力能效评价与标准化体系成为支撑绿色算力高质量发展的基础性研究课题。

（五）实验仿真与工程实践融合

算力能效优化的研究重心从纯算法突破转向工程体系的长期演进，实验仿真与产业实践的协同程度事关技术落地及节能成效。研究机构多聚焦调度算法、能耗预测模型、冷却策略优化[81,83]，产业界更关注模型可解释性、运行确定性、经济收益、安全可信等工程属性[91]。在能源与算力深度耦合的背景下，研究对象的层级也体现出差异性，即从单一算力系统的实验场景转向算力系统与电力系统协同运行的工程验证需求。这些结构性差异可能催生两类新的合作模式：面向算力与能源系统的联合实验平台，通过数字孪生方式对算力调度、电力供给、能碳耦合优化、冷却自适应方案等进行协同仿真与验证；科研云平台共享算力实践，如AWS、Azure、国内算力服务商开放能效软件开发工具包、节点遥测应用程序编程接口、节能算例库，支持研究模型在引入电价信号、碳因子、能源约束条件下从模拟环境应用走向真实生产系统[108]。国内外的监管部门和行业组织也在推动跨主体协作，要求算力节点报送能效状态，同步披露碳因子、能源结构等信息，为算力调度、电力调度的协同决策提供工程基础条件[3,12]。在此趋势下，对具有系统工程能力的复合型人才的需求持续攀升，研究方向拓展至能源计算、网络调度、功率电子、机房热工以及与AI系统的交叉，实验范式也从单一算法性能评测升级为涵盖节能效果、运行稳定性、能源协同、碳指标的综合验证体系，以稳健支撑面向算力与能源协同演化的长期工程实践。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: November 13, 2025; **Revised date:** December 30, 2025
Corresponding author: Yuan Yige is an associate research fellow from Xiangjiang Laboratory. Her major research fields inlude artificial intelligence, digital intelligent marketing. E-mail: immyyuan23@163.com
Funding project: National Natural Science Foundation of China (72088101); Chinese Academy of Engineering project “Strategic Research on the Construction of Digital Ecological Civilization Empowered by New-Generation Information Technology” (2023-JB-09); Xiangjiang Laboratory Project (24XJ01001, 23XJ01002)

参考文献

[1] 陈晓红, 曹廖滢, 陈姣龙, 等. 我国算力发展的需求、电力能耗及

绿色低碳转型对策 [J]. 中国科学院院刊, 2024, 39(3): 528–539.
Chen X H, Cao L Y, Chen J L, et al. Development demand, power energy consumption and green and low-carbon transition for computing power in China [J]. Bulletin of Chinese Academy of Sciences, 2024, 39(3): 528–539.

[2] 段晓东, 姚惠娟, 付月霞, 等. 面向算网一体化演进的算力网络技术 [J]. 电信科学, 2021, 37(10): 76–85.
Duan X D, Yao H J, Fu Y X, et al. Computing force network technologies for computing and network integration evolution [J]. Telecommunications Science, 2021, 37(10): 76–85.

[3] 管晓宏, 徐占伯, 吴江, 等. 数字基础设施绿色低碳发展中的关键科学问题与建议 [J]. 中国科学基金, 2024, 38(4): 583–592.
Guan X H, Xu Z B, Wu J, et al. Emerging topics in the green and low-carbon development of digital infrastructure [J]. Bulletin of National Natural Science Foundation of China, 2024, 38(4): 583–592.

[4] International Energy Agency. Data centres and data transmission networks: Energy system analysis [EB/OL] (2024-03-25)[2025-12-30]. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

[5] 算力基础设施高质量发展行动计划 [EB/OL]. (2023-10-08)[2025-12-30]. https://www.miit.gov.cn/zwgk/zcwj/wjfb/tz/art/2023/art_fcb3aa793e674960b1c00d7e3b6ad448.html.
Action plan for high-quality development of computing power infrastructure [EB/OL]. (2023-10-08)[2025-12-30]. https://www.miit.gov.cn/zwgk/zcwj/wjfb/tz/art/2023/art_fcb3aa793e674960b1c00d7e3b6ad448.html.

[6] Kou H. Power for AI: Easier said than built [EB/OL]. (2025-04-15)[2025-12-30]. https://about.bnef.com/insights/commodities/power-for-ai-easier-said-than-built/.

[7] Google Sustainability. 2023 environmental report [EB/OL]. (2023-07-15)[2025-12-30]. https://sustainability.google/reports/google-2023-environmental-report/.

[8] Alibaba cloud for sustainability [EB/OL]. (2023-06-15)[2025-12-30]. https://www.alibabacloud.com/solutions/sustainability.

[9] Shobanke M, Bhatt M, Shittu E. Advancements and future outlook of artificial intelligence in energy and climate change modeling [J]. Advances in Applied Energy, 2025, 17: 100211.

[10] Mahmood M, Chowdhury P, Yeassin R, et al. Impacts of digitalization on smart grids, renewable energy, and demand response: An updated review of current applications [J]. Energy Conversion and Management: X, 2024, 24: 100790.

[11] 郭远游, 叶玉瑶, 王长建, 等. “东数西算”战略背景下中国数据中心碳排放空间转移研究 [J]. 地理科学, 2025, 45(3): 459–471.
Guo Y Y, Ye Y Y, Wang C J, et al. Spatial transfer of carbon emissions in China’s data centers under background of “East Data and West Calculation” project [J]. Scientia Geographica Sinica, 2025, 45(3): 459–471.

[12] 苏晨晨, 王飞. 数字新基建与碳排放强度: 内在机制与经验证据 [J]. 地域研究与开发, 2025, 44(4): 36–43.
Su C C, Wang F. Digital new infrastructure and carbon emission intensity: Internal mechanisms and empirical evidence [J]. Areal Research and Development, 2025, 44(4): 36–43.

[13] Mytton D, Ashtine M. Sources of data center energy estimates: A comprehensive review [J]. Joule, 2022, 6(9): 2032–2056.

[14] Zhao P G, Lou B W, Luo Y, et al. High-powered computing: Energy consumption and its environmental impact [J]. Advances in Engineering Technology Research, 2025, 13(1): 1337.

[15] 邢文娟, 雷波, 赵倩颖. 算力基础设施发展现状与趋势展望 [J]. 电信科学, 2022, 38(6): 51–61.
Xing W J, Lei B, Zhao Q Y. Development status and trend prospect of computing power infrastructure [J]. Telecommunications Science, 2022, 38(6): 51–61.

[16] 李平, 邓洲, 张艳芳. 新科技革命和产业变革下全球算力竞争格局及中国对策 [J]. 经济纵横, 2021 (4): 33–42.
Li P, Deng Z, Zhang Y F. The global computing power competition pattern under the new sci-tech revolution and industrial transformation and China’s countermeasures [J]. Economic Review Journal, 2021 (4): 33–42.

[17] 张志学, 华中生, 谢小云. 数智时代人机协同的研究现状与未来方向 [J]. 管理工程学报, 2024, 38(1): 1–13.
Zhang Z X, Hua Z S, Xie X Y. Research status and future directions of human–computer collaboration in the era of digital intelligence [J]. Journal of Industrial Engineering and Engineering Management, 2024, 38(1): 1–13.

[18] Li Y L, Wen Y G, Tao D C, et al. Transforming cooling optimization for green data center via deep reinforcement learning [J]. IEEE Transactions on Cybernetics, 2020, 50(5): 2002–2013.

[19] Deenadayal T. Carbon emission reduction through AI-based energy optimization in data centers [J]. Global Journal of Engineering Innovations & Interdisciplinary Research, 2025, 5(3): 1–6.

[20] 陈志韬, 安虹, 邱晓杰, 等. 商用处理器上针对能耗优化的DVFS调节机制 [J]. 计算机工程, 2017, 43(3): 46–50, 56.
Chen Z T, An H, Qiu X J, et al. Energy consumption optimization DVFS mechanism for commercial processors [J]. Computer Engineering, 2017, 43(3): 46–50, 56.

[21] 陈敏, 高赐威, 郭庆来, 等. 互联网数据中心负荷时空可转移特性建模与协同优化: 驱动力与研究架构 [J]. 中国电机工程学报, 2022, 42(19): 6945–6958.
Chen M, Gao C W, Guo Q L, et al. Modeling and coordinated optimization for spatiotemporal load regulation potentials of Internet data centers: Motivation and architecture [J]. Proceedings of the CSEE, 2022, 42(19): 6945–6958.

[22] Huang N, Li X, Xu Q M, et al. Artificial intelligence-based temperature twinning and pre-control for data center airflow organization [J]. Energies, 2023, 16(16): 6063.

[23] 王丽莉, 赵飞龙. 基于风电升压站的低PUE数据中心实现 [J]. 电子测量技术, 2023, 46(18): 16–22.
Wang L L, Zhao F L. Implementation of low PUE data center based on offshore substation [J]. Electronic Measurement Technology, 2023, 46(18): 16–22.

[24] Li X Y, Zhou T, Wang H Y, et al. Energy-efficient computation with DVFS using deep reinforcement learning for multi-task systems in edge computing [J]. IEEE Transactions on Sustainable Computing, 2025, 10(6): 1116–1127.

[25] 杨挺, 姜含, 侯昱丞, 等. 基于计算负荷时–空双维迁移的互联多数据中心碳中和调控方法研究 [J]. 中国电机工程学报, 2022, 42(1): 164–176.
Yang T, Jiang H, Hou Y C, et al. Study on carbon neutrality regu-

lation method of interconnected multi-datacenter based on spatio-temporal dual-dimensional computing load migration [J]. Proceedings of the CSEE, 2022, 42(1): 164–176.

[26] Chen G Z, Lu S Z, Zhou S Y, et al. A systematic review of building energy consumption prediction: From perspectives of load classification, data-driven frameworks, and future directions [J]. Applied Sciences, 2025, 15(6): 3086.

[27] 李远征, 龙信鑫, 周纯杰, 等. 高比例新能源电网–分布式数据中心集群协同优化运行研究 [J]. 中国科学: 技术科学, 2024, 54 (1): 119–135.
Li Y Z, Long X X, Zhou C J, et al. Coordinated operations of highly renewable power systems and distributed data centers [J]. Scientia Sinica Technologica, 2024, 54(1): 119–135.

[28] Dayarathna M, Wen Y G, Fan R. Data center energy consumption modeling: A survey [J]. IEEE Communications Surveys & Tutorials, 2016, 18(1): 732–794.

[29] Latif I, Newkirk A C, Carbone M R, et al. Single-node power demand during AI training: Measurements on an 8-GPU NVIDIA H100 system [J]. IEEE Access, 2025, 13: 61740–61747.

[30] Chen H W, Li D. Current status and challenges for liquid-cooled data centers [J]. Frontiers in Energy Research, 2022, 10: 952680.

[31] Xu S J, Zhang H, Wang Z L, et al. Thermal management and energy consumption in air, liquid, and free cooling systems for data centers: A review [J]. Energies, 2023, 16(3): 1279.

[32] Desislavov R, Martínez-Plumed F, Hernández-Orallo J. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning [J]. Sustainable Computing: Informatics and Systems, 2023, 38: 100857.

[33] Kanbur B B, Wu C L, Fan S M, et al. Two-phase liquid-immersion data center cooling system: Experimental performance and thermoeconomic analysis [J]. International Journal of Refrigeration, 2020, 118: 290–301.

[34] Yu J, Kim J, Seo E. Know your enemy to save cloud energy: Energy-performance characterization of machine learning serving [R]. Montreal: 2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2023.

[35] Jaureguialzo E. PUE: The green grid metric for evaluating the energy efficiency in dc (data center). measurement method using the power demand [R]. Amsterdam: 2011 IEEE 33rd International Telecommunications Energy Conference, 2011.

[36] Ling C L, Tang J Q, Zhao P J, et al. Unraveling the relation between carbon emission and carbon footprint: A literature review and framework for sustainable transportation [J]. npj Sustainable Mobility and Transport, 2024, 1: 13.

[37] Liu Z, Qiu Z W. A systematic review of transportation carbon emissions based on CiteSpace [J]. Environmental Science and Pollution Research, 2023, 30(19): 54362–54384.

[38] Ma Z G, Sun T. Study on measurement and driving factorsof carbon emission intensity from energy consumption in China [J]. Polish Journal of Environmental Studies, 2022, 31(4): 3687–3699.

[39] Liu X O. Research on collaborative scheduling of Internet data center and regional integrated energy system based on electricity-heat-water coupling [J]. Energy, 2024, 292: 130462.

[40] Lei N A, Masanet E. Climate- and technology-specific PUE and WUE estimations for U.S. data centers using a hybrid statistical and thermodynamics-based approach [J]. Resources, Conservation and Recycling, 2022, 182: 106323.

[41] Asperti A, Evangelista D, Marzolla M. Dissecting flops along input dimensions for greenai cost estimations [EB/OL]. (2021-07-26)[2025-12-30]. https://arxiv.org/abs/2107.11949.

[42] 张立博, 李昌伟, 齐伟, 等. 神经网络训练处理器的浮点运算优化架构 [J]. 计算机测量与控制, 2023, 31(6): 176–182.
Zhang L B, Li C W, Qi W, et al. Floating point optimization architecture of neural network training processor [J]. Computer Measurement & Control, 2023, 31(6): 176–182.

[43] Li C, Tsourdos A, Guo W S. A transistor operations model for deep learning energy consumption scaling law [J]. IEEE Transactions on Artificial Intelligence, 2024, 5(1): 192–204.

[44] Yu G C, Wang Z Y, Xu Y L, et al. From energy to ecology: Decarbonization pathways for sustainable high-performance computing through global carbon-energy nexus analysis [J]. Frontiers in Applied Mathematics and Statistics, 2025, 11: 1595365.

[45] Li X Y, Zhao Z M, Jiang X J, et al. A review on AI-driven optimization of data center energy efficiency and thermal management [J]. International Journal of Applied Science, 2025, 8(3): 108.

[46] Ratkovic I, Bezanic N, Ünsal O S, et al. An overview of architecture-level power- and energy-efficient design techniques [J]. Advances in Computers, 2015, 98: 1–57.

[47] Cong T, Charot F. Design space exploration of heterogeneous-accelerator SoCs with hyperparameter optimization [R]. Tokyo: The 26th Asia and South Pacific Design Automation Conference, 2021.

[48] Belkin S. Air-immersion solution, maximizing data centers heat reuse, using hybrid cooling approach combining two-phase direct on chip dielectric liquid and air-based chip cooling [R]. San Jose : 2023 39th Semiconductor Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), 2023.

[49] Hnayno M, Chehade A, Klaba H, et al. Experimental investigation of a data-centre cooling system using a new single-phase immersion/liquid technique [J]. Case Studies in Thermal Engineering, 2023, 45: 102925.

[50] Miao Z C, Liu L, Nan H J, et al. Energy and carbon-aware distributed machine learning tasks scheduling scheme for the multi-renewable energy-based edge-cloud continuum [J]. Science and Technology for Energy Transition, 2024, 79: 82.

[51] Zhong Z H, Buyya R. A cost-efficient container orchestration strategy in Kubernetes-based cloud computing infrastructures with heterogeneous resources [J]. ACM Transactions on Internet Technology, 2020, 20(2): 1–24.

[52] Duan H C, Chen C, Min G Y, et al. Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems [J]. Future Generation Computer Systems, 2017, 74: 142–150.

[53] Patel P D. Artificial intelligence in datacenters: Optimizing performance, power, and thermal management [J]. Journal of Computer Science and Technology Studies, 2025, 7(4): 952–963.

[54] Fulpagare Y, Huang K R, Liao Y H, et al. Optimal energy management for air cooled server fans using deep reinforcement learning control method [J]. Energy and Buildings, 2022, 277: 112542.

[55] Goel S, Bajpai M. AI-driven energy management in green cloud computing: A systematic review [J]. International Journal for Multidisciplinary Research, 2025, 7(3): 48659.

[56] Sreekumar G, Martin J P, Raghavan S, et al. Transformer-based forecasting for sustainable energy consumption toward improving socioeconomic living: AI-enabled energy consumption forecasting [J]. IEEE Systems, Man, and Cybernetics Magazine, 2024, 10 (2): 52–60.

[57] Ardabili S, Abdolalizadeh L, Mako C, et al. Systematic review of deep learning and machine learning for building energy [J]. Frontiers in Energy Research, 2022, 10: 786027.

[58] Zou S L, Luo X J, Yang Z X. Energy consumption forecasting in buildings based on long- term and short-term memory networks [R]. Melbourne: 2024 2nd International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII), 2024.

[59] Yang Z Y, Gaidhane A D, Drgoňa J, et al. Physics-constrained graph modeling for building thermal dynamics [J]. Energy and AI, 2024, 16: 100346.

[60] Tashiro S, Nakamura Y, Matsuda K, et al. Application of convolutional neural network to prediction of temperature distribution in data centers [R]. San Francisco: 2016 IEEE 9th International Conference on Cloud Computing (CLOUD), 2016.

[61] Chen C, Mithani N, Jin T, et al. Liquid cooling practice on meta's AI training platform [R]. Garden Grove: ASME 2022 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems, 2022.

[62] Cui L Z, Xu C, Yang S, et al. Joint optimization of energy consumption and latency in mobile edge computing for Internet of things [J]. IEEE Internet of Things Journal, 2019, 6(3): 4791–4803.

[63] 梁晨, 曾博, 雷乐意, 等. 算力–电力联合市场下数据中心与配电网集成规划: 一种多目标区间–随机优化方法 [J]. 电力系统保护与控制, 2025, 53(16): 120–135.
Liang C, Zeng B, Lei L Y, et al. Integrated planning of data centers and distribution networks under the computing-electricity joint market: A multi-objective interval-stochastic optimization approach [J]. Power System Protection and Control, 2025, 53(16): 120–135.

[64] Anusha P, Balan R V. Efficient power management in mobile computing with edge server offloading using multi-objective optimization [J]. EAI Endorsed Transactions on Energy Web, 2018: 170288.

[65] 张维庭, 任家栋, 张宏科. 算力网络架构的演进与创新 [J]. 科技导报, 2025, 43(9): 24–30.
Zhang W T, Ren J D, Zhang H K. Evolution and innovative paradigms of computing–aware networks architecture [J]. Science & Technology Review, 2025, 43(9): 24–30.

[66] Yi D L, Zhou X, Wen Y G, et al. Efficient compute-intensive job allocation in data centers via deep reinforcement learning [J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 31 (6): 1474–1485.

[67] Peng Y, Shen H J, Tang X C, et al. Energy consumption optimization for heating, ventilation and air conditioning systems based on deep reinforcement learning [J]. IEEE Access, 2023, 11: 88265–88277.

[68] Chen Y, Ruan P, Wang G. Research on the application effects of AI technology in data center energy saving [R]. Chengdu: The 2024 2nd International Conference on Electronics, Computers and Communication Technology, 2024:

[69] Hu J R, Liu L, Liu S, et al. Co-optimization of GPU AI chip from technology, design, system and algorithms [R]. San Francisco: 2024 IEEE International Electron Devices Meeting (IEDM), 2024.

[70] Zhang J, Rangineni K, Ghodsi Z, et al. Thundervolt: Enabling aggressive voltage underscaling and timing error resilience for energy efficient deep learning accelerators [R]. San Francisco: The 55th Annual Design Automation Conference, 2018.

[71] Jin R. Deep learning at Alibaba [R]. Melbourne: The Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.

[72] Ji Z L, Qin Z J, Tao X M. Meta federated reinforcement learning for distributed resource allocation [J]. IEEE Transactions on Wireless Communications, 2024, 23(7): 7865–7876.

[73] Mu N, Luan Y, Jia Q S. Preference-based multi-objective reinforcement learning [J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 18737–18749.

[74] Liu C M, Xu X, Hu D W. Multiobjective reinforcement learning: A comprehensive overview [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2015, 45(3): 385–398.

[75] Siddique I. Carbon-aware cloud computing: AI-driven predictive modeling and dynamic optimization of data center energy consumption and emission reduction strategies [J]. Physical Education, Health and Social Sciences, 2025, 3(3): 41–55.

[76] Tanash M, Dunn B, Andresen D, et al. Improving HPC system performance by predicting job resources via supervised machine learning [R]. Chicago: The Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning), 2019.

[77] 李传江, 李少波, 冯毅雄, 等. "东数西算"驱动的工业5.0算力网络: 创新架构与协同应用 [J]. 计算机集成制造系统, 2025, 31 (10): 3567–3581.
Li C J, Li S B, Feng Y X, et al. Industrial 5.0 computing power network driven by "Eastern Data and Western Computing": Innovative architecture and collaborative applications [J]. Computer Integrated Manufacturing Systems, 2025, 31(10): 3567–3581.

[78] 金驰, 潘京津. 绿色数据中心评价指标体系解读 [J]. 信息技术与标准化, 2020 (12): 29–33.
Jin C, Pan J J. Interpretation of the evaluation index system of green data center [J]. Information Technology & Standardization, 2020 (12): 29–33.

[79] Thum K E, Shasha D E, Lejay L V, et al. Light- and carbon-signaling pathways modeling circuits of interactions [J]. Plant Physiology, 2003, 132(2): 440–452.

[80] 陈启鑫, 康重庆, 夏清, 等. 低碳电力调度方式及其决策模型 [J]. 电力系统自动化, 2010, 34(12): 18–23.
Chen Q X, Kang C Q, Xia Q, et al. Mechanism and modelling approach to low-carbon power dispatch [J]. Automation of Electric Power Systems, 2010, 34(12): 18–23.

[81] 顾健华, 冯建华, 许辉阳, 等. 基于有向图与卷积网络强化学习的端侧协同算力资源分配方法 [J]. 电子学报, 2025, 53(6): 1771–1783.

Gu J H, Feng J H, Xu H Y, et al. Directed graph and convolutional network reinforcement learning for terminal-side collaborative computing resource allocation scheme [J]. Acta Electronica Sinica, 2025, 53(6): 1771–1783.

[82] 胡春潮. 分布式智能电网自动化关键技术研究与工程应用 [D]. 广州: 华南理工大学(博士学位论文), 2022.
Hu C C. Research on key technologies and engineering applications of distributed smart grid automation [D]. Guangzhou: South China University of Technology (Doctoral dissertation), 2022.

[83] 李学龙. 多模态认知计算 [J]. 中国科学: 信息科学, 2023, 53(1): 1–32.
Li X L. Multi-modal cognitive computing [J]. Scientia Sinica Informationis, 2023, 53(1): 1–32.

[84] Hob M, Kranzlmuller D. Enable energy efficient data centers [R]. Singapore: The HPC Asia 2023 Workshops, 2023.

[85] Bhatt U, Xiang A, Sharma S, et al. Explainable machine learning in deployment [R]. Barcelona: The 2020 Conference on Fairness, Accountability, and Transparency, 2020.

[86] 穆旭彤, 程珂, 宋安霄, 等. 抗拜占庭攻击的隐私保护联邦学习 [J]. 计算机学报, 2024, 47(4): 842–861.
Mu X T, Cheng K, Song A X, et al. Privacy-preserving federated learning resistant to Byzantine attacks [J]. Chinese Journal of Computers, 2024, 47(4): 842–861.

[87] 丁煜蓉, 陈红坤, 吴军, 等. 计及综合能效的电–气–热综合能源系统多目标优化调度 [J]. 电力系统自动化, 2021, 45(2): 64–73.
Ding Y R, Chen H K, Wu J, et al. Multi-objective optimal dispatch of electricity–gas–heat integrated energy system considering comprehensive energy efficiency [J]. Automation of Electric Power Systems, 2021, 45(2): 64–73.

[88] 刘炎培, 朱运静, 宾艳茹, 等. 边缘环境下计算密集型任务调度研究综述 [J]. 计算机工程与应用, 2022, 58(20): 28–42.
Liu Y P, Zhu Y J, Bin Y R, et al. Review of research on computing-intensive task scheduling in edge environments [J]. Computer Engineering and Applications, 2022, 58(20): 28–42.

[89] 段晓东, 刘鹏, 陆璐, 等. 确定性网络技术综述 [J]. 电信科学, 2023, 39(11): 1–12.
Duan X D, Liu P, Lu L, et al. Review of deterministic networking technology [J]. Telecommunications Science, 2023, 39(11): 1–12.

[90] 李清清, 于欣宁, 王海峰. GPU异构集群的协同计算引擎设计研究 [J]. 计算机应用与软件, 2024, 41(12): 15–22, 28.
Li Q Q, Yu X N, Wang H F. Collaborative computing engine design for GPU heterogeneous cluster [J]. Computer Applications and Software, 2024, 41(12): 15–22, 28.

[91] 郑文旭, 潘晓东, 马迪, 等. 用于高性能计算的作业调度能效性研究综述 [J]. 计算机工程与科学, 2019, 41(9): 1525–1533.
Zheng W X, Pan X D, Ma D, et al. Overview on the energy efficiency of job scheduling for high performance computing [J]. Computer Engineering & Science, 2019, 41(9): 1525–1533.

[92] 任杰, 高岭, 于佳龙, 等. 面向边缘设备的高能效深度学习任务调度策略 [J]. 计算机学报, 2020, 43(3): 440–452.
Ren J, Gao L, Yu J L, et al. Energy-efficient deep learning task scheduling strategy for edge device [J]. Chinese Journal of Computers, 2020, 43(3): 440–452.

[93] 刘虎沉, 王鹤鸣, 施华. 智能质量管理: 理论模型、关键技术与研究展望 [J]. 中国管理科学, 2024, 32(3): 287–298.
Liu H C, Wang H M, Shi H. Intelligent quality management: Theoretical framework, key technologies, and research prospect [J]. Chinese Journal of Management Science, 2024, 32(3): 287–298.

[94] 李萌, 黄钰典, 杨睿哲, 等. 面向绿色低碳的工业互联网: 发展与挑战 [J]. 北京工业大学学报, 2023, 49(11): 1251–1262.
Li M, Huang Y D, Yang R Z, et al. Industrial Internet for green and low carbon: Developments and challenges [J]. Journal of Beijing University of Technology, 2023, 49(11): 1251–1262.

[95] 张川, 胡沛裕, 殷格格, 等. 低碳能源系统中能源利用技术现状及展望 [J]. 中国工程科学, 2024, 26(4): 164–175.
Zhang C, Hu P Y, Yin G G, et al. Comprehensive review and future trend outlook on energy utilization technologies in low-carbon energy systems [J]. Strategic Study of CAE, 2024, 26(4): 164–175.

[96] 倪静, 王振全, 易久, 等. 绿色数据中心能耗评价指标体系研究 [J]. 电气应用, 2014, 33(8): 89–93.
Ni J, Wang Z Q, Yi J, et al. Research on energy consumption evaluation index system of green data center [J]. Electrotechnical Application, 2014, 33(8): 89–93.

[97] 徐丹, 曾宇, 孟维业, 等. AI使能的5G节能技术 [J]. 电信科学, 2021, 37(5): 32–41.
Xu D, Zeng Y, Meng W Y, et al. AI-enabled 5G energy-saving technology [J]. Telecommunications Science, 2021, 37(5): 32–41.

[98] 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 [J]. 中国科学: 信息科学, 2020, 50(1): 1–24.
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 [J]. Scientia Sinica Informationis, 2020, 50(1): 1–24.

[99] 唐文虎, 陈星宇, 钱曈, 等. 面向智慧能源系统的数字孪生技术及其应用 [J]. 中国工程科学, 2020, 22(4): 74–85.
Tang W H, Chen X Y, Qian T, et al. Technologies and applications of digital twin for developing smart energy systems [J]. Strategic Study of CAE, 2020, 22(4): 74–85.

[100] Hamdan A, Ibekwe K I, Ilojianya V I, et al. AI in renewable energy: A review of predictive maintenance and energy optimization [J]. International Journal of Science and Research Archive, 2024, 11(1): 718–729.

[101] Ukoba K, Olatunji K O, Adeoye E, et al. Optimizing renewable energy systems through artificial intelligence: Review and future prospects [J]. Energy & Environment, 2024, 35(7): 3833–3879.

[102] Kumar A, He X N, Deng Y, et al. Earth grid: Toward a low-carbon energy infrastructure [J]. iScience, 2025, 28(11): 113681.

[103] Li T, Hou J, Yan J L, et al. Chiplet heterogeneous integration technology—Status and challenges [J]. Electronics, 2020, 9(4): 670.

[104] Shan G B, Zheng Y W, Xing C Y, et al. Architecture of computing system based on chiplet [J]. Micromachines, 2022, 13(2): 205.

[105] 邢文娟, 雷波, 赵倩颖. 算力基础设施发展现状与趋势展望 [J]. 电信科学, 2022, 38(6): 51–61.
Xing W J, Lei B, Zhao Q Y. Development status and trend prospect of computing power infrastructure [J]. Telecommunications Science, 2022, 38(6): 51–61.

[106] Wojtaszek H. Energy transition 2024—2025: New demand vec-

tors, technology oversupply, and shrinking net-zero 2050 premium [J]. Energies, 2025, 18(16): 4441.

[107] Bhandari K P, Collier J M, Ellingson R J, et al. Energy payback time (EPBT) and energy return on energy invested (EROI) of solar photovoltaic systems: A systematic review and meta-analysis [J]. Renewable and Sustainable Energy Reviews, 2015, 47: 133–141.

[108] Palumbo F, Aceto G, Botta A, et al. Characterization and analysis of cloud-to-user latency: The case of Azure and AWS [J]. Computer Networks, 2021, 184: 107693.