基于仿真模型的数据中心 PUE 协同优化方法及工程验证

胡 涛

（中国电信长三角国家枢纽算力中心，浙江 嘉兴 314100）

摘要：针对当前数据中心能源管理粗放化、电能使用效率（PUE）偏高、制冷系统与 IT 负载动态协同能力不足，以及直接工程改造风险高、效益难以量化评估等问题，本文提出了一种基于数字仿真模型的 PUE 协同优化与验证方法。该方法构建了“仿真建模－算法寻优－策略验证－工程部署”的全流程技术体系。首先，通过分析数据中心能耗构成及 PUE 关键影响因素，基于历史运行数据搭建了融合 IT 负载预测、制冷系统动态调节及计算负载智能分配的一体化高保真仿真环境。在此基础上，融合双向长短期记忆网络时序预测与深度确定性策略梯度强化学习算法，构建了多目标协同优化决策模型，并通过多场景仿真确定了最优控制策略。最后，将优化策略部署于某中型数据中心进行实地工程验证。仿真结果表明，优化后系统平均 PUE 由 1.48 降至 1.32，制冷系统能耗降低 22.4%，服务器集群负载均衡度提升 26.4%；为期 6 个月的工程验证显示，实际运行年均 PUE 稳定在 1.32，年节约电费约 82 万元，静态投资回收期约为 1.2 年。本方法通过仿真前置有效规避了盲目改造风险，为数据中心绿色低碳运行提供了可量化、可复制的技术路径。

关键词：数据中心；电能使用效率优化；数字仿真；时序预测；强化学习；能耗协同控制；节能

收稿日期：2026 年 2 月 26 日 **中图分类号：**TN876.5

通讯作者：胡涛 中国电信长三角国家枢纽算力中心

Collaborative optimization method and engineering verification of pue in data center based on simulation model

Hu Tao

（Chinatelecom Yangtze River Delta national hub Computing Center, Zhejiang Jiaxing 314100）

Abstract: In view of the problems such as extensive energy management, high Power Usage Effectiveness（PUE）, insufficient dynamic coordination between cooling systems and IT loads, high risks associated with direct engineering retrofits, and difficulty in quantifying the benefits in current data centers, this paper proposes a PUE collaborative optimization and verification method based on a digital simulation model. This method establishes a full-process technical system of "simulation modeling – algorithm optimization – strategy verification – engineering deployment". Firstly, by analyzing the energy consumption composition and key influencing factors of PUE in data centers, a high-fidelity integrated simulation environment incorporating IT load forecasting, dynamic adjustment of cooling systems, and intelligent distribution of computational loads is constructed based on historical operational data. On this basis, an optimization decision-making model for multi-objective collaboration is built by integrating a Bidirectional Long Short-Term Memory network for time series prediction and a Deep Deterministic Policy Gradient reinforcement learning algorithm. The optimal control strategy is determined through multi-scenario simulations. Finally, the

optimized strategy is deployed in a medium-sized data center for on-site engineering validation. Simulation results show that the average PUE decreases from 1.48 to 1.32, the energy consumption of the cooling system is reduced by 22.4%, and the load balance of the server cluster is improved by 26.4%. A six-month engineering validation demonstrates that the actual annual average PUE remains stable at 1.32, achieving an annual electricity cost saving of approximately 820,000 yuan with a static investment payback period of about 1.2 years. By prioritizing simulation, this method effectively mitigates the risks associated with blind retrofits and provides a quantifiable, replicable technical pathway for the green and low-carbon operation of data centers.

Keywords: Data Center; PUE Optimization; Digital Simulation; Time Series Prediction; Reinforcement Learning; Collaborative Energy Consumption Control; Energy Saving

0 引言

随着数字经济高速发展，云计算、人工智能及大数据等技术的规模化应用持续驱动数据中心建设规模与算力需求的爆发式增长[1]。行业数据显示，我国数据中心机架总规模近年来保持年均 30% 以上的增速，截至 2024 年底，在用标准机架数已超过 300 万，总算力规模突破 200 EFLOPS。与此同时，数据中心的高能耗问题日益凸显，其年耗电量已超过 2000 亿千瓦时，约占全国用电总量的 2.5%，且年均增长率维持在 15% ~ 20% 的高位。电能使用效率（PUE）作为衡量数据中心能源效率的核心指标[17]，我国多数存量数据中心的 PUE 值仍维持在 1.45 以上，与国际先进水平（≤ 1.3）存在较大差距，节能降耗潜力巨大。

现有的 PUE 优化方法主要可分为两类：一是以更新高效制冷设备、改进气流组织等为主的硬件改造型方案[4][10]，其往往面临投资成本高、改造周期长、实施风险难以预估等挑战；二是以调节运行参数为主的软件调控型方案[2][8]，但现有方法多聚焦于制冷或 IT 系统等单一维度的优化，缺乏跨系统的协同控制，且优化效果的预评估手段不足。数字仿真建模技术能够以较低成本构建反映数据中心全系统运行特性的虚拟环境[6]，为优化算法的验证与效果预判提供了有效工具。

鉴于此，本文贯彻“仿真先行，精准落地”的理念，旨在构建一个数据中心一体化仿真模型，通过融合时序预测与强化学习算法[3][7]，实现 IT 负载、制冷系统、供电系统的协同动态优化。本研究通过详尽的仿真分析与实地工程验证，形成了一套完整的方法论与实践体系，旨在为数据中心，尤其是存量数据中心的绿色节能改造与精细化管理提供可靠的技术支撑[9][13]。

1 面向 PUE 优化的数据中心仿真建模

构建能够精准反映数据中心实际运行状态与动态特性的仿真环境，是进行算法验证与优化策略寻优的基础。本节从关键影响因素剖析、仿真框架设计及环境参数配置三个方面阐述建模过程。

1.1 PUE 关键影响因素分析

数据中心总能耗主要由 IT 设备能耗与基础设施能耗构成。其中，IT 设备能耗与服务器负载率、硬件能效直接相关；基础设施能耗中，制冷系统占比通常超过 60%，是 PUE 优化的主要对象[12][19]。基于对某中型数据中心连续 3 个月的运行数据分析，提炼出影响 PUE 的四大类核心因素：

（1）IT 负载特性：服务器集群的负载率及其在机架间的分布均匀性[16][20]。通常负载率在 60%-80% 区间能效较优，负载波动剧烈或分布不均会导致局部热点，增加不必要的制冷能耗。

（2）制冷系统运行状态：包括冷冻水供水温度、送回风温度、风机频率、冷水机组运行台数与负载率、设备性能系数（COP/EER）等[8][15]。制冷量与热负荷的精确匹配是节能关键。

（3）供配电系统效率：变压器、不间断电源（UPS）的负载率与转换效率。变压器在 50% ~ 70% 负载率、UPS 在 92% ~ 96% 负载区间通常效率最高，线路损耗亦不容忽视。

（4）外部环境条件：室外温湿度直接影响自然冷却时长与冷机能效；机房围护结构保温性能影响冷量损失[18]。上述因素相互耦合，建模需充分考虑其协同影响机制。

1.2 仿真模型总体框架设计

设计“数据层 - 模型层 - 控制层 - 评估层”

四层仿真框架（如图 1 所示）：

（1）数据层：负责数据采集与预处理。采集涵盖 12 个月的历史运行数据（约 17.5 万条）、实时工况数据及设备铭牌参数。采用 3σ 准则处理异常值，均值填充与相似日法补齐缺失值，并进行 Min-Max 标准化，形成高质量数据集。

（2）模型层：仿真核心，包含设备模型与核心算法模型。设备模型涵盖服务器、冷水机组、空调末端、泵组、变压器、UPS 等，基于物理原理与数据驱动方法构建[6][20]。核心算法模型包括 IT 负载与温度预测模型、PUE 优化决策模型。

（3）控制层：采用分布式控制架构模拟实际控制系统，分为设备级控制（如调节风机频率）和系统级协同控制（如根据负载预测调节冷机出水温度），并集成计算流体动力学（CFD）简化模型模拟气流组织与热传递[4][10]。

（4）评估层：实时计算并输出 PUE、局部 PUE、制冷能耗占比、负载均衡度、温度标准差等关键性能指标（KPI）[17]，支持周期性评估与对比分析，为策略调优提供量化依据。

1.3 仿真环境参数配置

基于 Python 与 TensorFlow 搭建仿真平台，具体配置如下：

（1）硬件平台：Intel i9-12900K CPU，NVIDIA RTX 3090 GPU，64GB RAM，存储为 1TB SSD + 4TB HDD，确保大规模并行计算与数据处理效率。

（2）软件环境：操作系统为 Windows Server 2022，编程语言为 Python 3.9，主要依赖库包括 TensorFlow 2.8, Pandas, NumPy 等，通过 Anaconda 管理虚拟环境[3][7]。

（3）场景参数：参考目标验证数据中心，设定仿真场景：机房面积 1200 ㎡，部署 280 台标准机架式服务器，制冷系统包含 3 台冷水机组及相应的冷却塔、水泵、空调末端。仿真初始化平均 PUE 为 1.48[5]。

（4）精度保障：通过历史数据对设备模型参数进行校准，确保稳态误差≤ 5%；仿真步长设为 15 分钟；合理设置边界条件（如室外气象参数），以保障仿真环境的真实性与准确性[6][11]。

2 融合预测与强化的 PUE 协同优化算法设计

为克服传统方法调节滞后与缺乏协同的局限，本节设计了一种融合时序预测与强化学习的协同优化算法，实现前瞻性决策与动态优化[2][12]。

2.1 基于 Bi-LSTM 的负载 - 温度耦合预测模型

针对 IT 负载与机房环境温度强耦合的特点，构建双向长短期记忆网络（Bi-LSTM）耦合预测模型，实现未来 1 小时负载与温度的双变量精准预测。

（1）模型结构：输入层为过去 24 小时的历史序列，特征包括 IT 总负载、机房平均温度、室外温湿度、时间标签等 5 类共 96 个数据点。隐藏层包含两个堆叠的 Bi-LSTM 层，神经元数分别为 64 和 32，用于捕捉序列的前后依赖关系。随后通过全连接融合层，输出层直接预测未来 1 小时的 IT 负载与机房平均温度。

（2）模型训练：采用 80% 的数据作为训练集，使用 Adam 优化器（初始学习率 0.001），以均方误差（MSE）作为损失函数，共进行 100 个训练轮次（epoch）[7][16]。

（3）性能验证：如表 1 所示，本模型在独立测试集上的表现优于传统 ARIMA 模型与单向 LSTM 模型。IT 负载预测的平均绝对误差（MAE）为 2.3%，均方根误差（RMSE）为 3.1%；温度预测的 MAE 为 0.4℃，RMSE 为 0.5℃。更高的预测精度为后续优化提供了可靠的前瞻信息[3][18]。

表 1 不同预测模型性能对比

预测模型	IT 负载预测 MAE（%）	IT 负载预测 RMSE（%）	温度预测 MAE（° C）	温度预测 RMSE（° C）
ARIMA 模型	4.2	5.7	0.67	0.83
单向 LSTM 模型	2.8	3.8	0.47	0.59
本文 Bi-LSTM 耦合模型	2.3	3.1	0.4	0.5

2.2 基于 DDPG 的 PUE 优化决策模型

将 PUE 优化问题建模为马尔可夫决策过程（MDP），采用深度确定性策略梯度（DDPG）算法进行求解，实现多约束下的连续动作空间优化[2][7]

[12]。

（1）MDP 建模：

状态空间 (S)：包含 8 类核心变量，如当前 IT 总负载、机房各区域温度、冷机出水温度、室外温湿度、当前 PUE 值等[8][17]。

动作空间 (A)：定义为 6 类可连续调节的控制指令，例如：冷水机组设定温度调整量、冷冻水泵频率调整量、虚拟负载迁移策略（引导计算任务至特定机架）等[14][16]。

奖励函数 (R)：设计为多目标加权和形式：R = −（ω1*PUE + ω2*P_cooling + ω3*T_variance + ω4*(1−L_balance)）。其中，P_cooling 为制冷功耗，T_variance 为机房温度标准差，L_balance 为负载均衡度。通过调整权重系数（本文设 ω1=0.5, ω2=0.25, ω3=0.15, ω4=0.1），引导智能体在降低 PUE、节约制冷耗电、保障温度稳定和提升负载均衡间取得平衡[19][20]。

（2）算法实现与训练：采用 Actor−Critic 双网络结构，并引入目标网络与经验回放池以提升训练稳定性。训练过程分为探索阶段和收敛阶段，当仿真环境中连续多个周期的平均 PUE 稳定≤ 1.35 时，判定模型收敛，停止训练。

2.3 仿真优化闭环流程设计

优化过程遵循以下六步闭环流程（如图 2 所示）：

（1）数据输入与初始化：载入预处理后的实时 / 历史数据，初始化仿真环境状态[6]。

（2）前瞻预测：调用训练好的 Bi−LSTM 预测模型，输出未来 1 小时的负载与温度预测值［3］。

（3）智能决策：DDPG 智能体根据当前状态及预测信息，输出最优控制动作集合[2][7]。

（4）仿真执行：仿真环境执行控制动作，驱动各设备模型运行，更新下一时刻的系统状态（包括能耗、温度等）[4][10]。

（5）效果评估与记录：评估层计算新状态下的 PUE 等指标，同时将本次状态转移（s, a, r, s’）存入经验池，用于 DDPG 模型更新[12][17]。

（6）迭代调优：循环执行步骤 2–5，直至达到预设的优化目标（如 PUE ≤ 1.32）或仿真周期结束。

3 仿真实验结果与分析

在搭建的仿真环境中，设置对照组（采用基于固定规则的现行策略）与实验组（采用本文提出的协同优化策略），进行为期 30 天（仿真时间）的对比实验，并从预测精度、整体能效及多场景适应性三个方面进行分析[6][11]。

3.1 预测模型精度验证

如第三部分表 1 所示，本文提出的 Bi−LSTM 耦合预测模型在各项指标上均显著优于对比模型。相较于 ARIMA 模型，其 IT 负载与温度预测的 MAE 分别降低了 45.2% 和 40.3%；相较于单向 LSTM 模型，分别降低了 17.9% 和 14.9%。这证明该模型能有效捕捉负载与温度间的复杂非线性关系及时序动态，为优化决策提供了高质量的前置输入[3][20]。

3.2 PUE 优化效果分析

对比实验组与对照组的核心能效指标，结果如表 2 所示。实验组的平均 PUE 由 1.48 降至 1.32，优化幅度达 10.8%。制冷系统平均能耗从 336 kW 降低至 260.3 kW，降幅达 22.4%，这是 PUE 降低的主要贡献来源[12][13]。同时，服务器集群的负载均衡度从 0.72 提升至 0.91，优化了 26.4%，表明算法有效引导了负载分布，避免了局部过热。机房温度波动幅度（标准差）从 2.1℃缩小至 1.3℃，热环境更加稳定[15][18]。在整个优化过程中，IT 设备总能耗未发生显著增长，实现了节能与算力保障的平衡[2][9]。

表 2 对照组与实验组仿真能效指标对比

评估指标	对照组	实验组	优化幅度 / 变化
平均 PUE	1.48	1.32	−10.8%
制冷系统平均能耗（kW）	336	260.3	−22.4%
服务器负载均衡度	0.72	0.91	+26.4%
机房温度波动标准差（° C）	2.1	1.3	−38.1%

3.3 多场景适应性验证

为检验算法在不同工况下的鲁棒性，设计了

三种典型场景进行仿真验证[8][11][18]：

（1）低负载场景（平均负载率≤ 40%）：优化前 PUE 较高（1.62），因制冷系统存在“大马拉小车”现象。优化后，算法通过提高冷机出水温度、减少泵频等策略，将 PUE 降至 1.45[12][15]。

（2）高负载场景（平均负载率≥ 80%）：优化前 PUE 为 1.42。优化后，算法在保障制冷需求的前提下，通过精细化送风控制和负载均衡，将 PUE 降至 1.28[16][19]。

（3）极端环境场景（模拟夏季高温高湿）：室外环境恶化导致制冷效率下降，优化前 PUE 为 1.55。优化后，算法动态调整运行策略，PUE 降至 1.39[18][20]。

如表 3 所示，在三种差异显著的场景下，本文方法均能实现 PUE 显著降低（≥ 10%）、制冷能耗有效削减（≥ 20%），并维持良好的负载均衡（≥ 0.89）与温度稳定性，证明了其广泛的适应性和有效性[7][12]。

表 3 多场景仿真验证结果

仿真场景	优化前平均PUE	优化后平均PUE	制冷能耗降低率（%）	负载均衡度
低负载场景	1.62	1.45	24.1	0.90
高负载场景	1.42	1.28	20.3	0.92
极端环境场景	1.55	1.39	21.7	0.89

4 工程实践与效果验证

为验证仿真优化策略的工程实用性，将训练得到的最优控制策略部署于某中型数据中心（参数与仿真场景一致）的实际能源管理系统中，进行了为期 6 个月的现场运行验证[6][9][11]。

4.1 部署与实施

在保证业务连续性的前提下，分阶段部署优化系统[14][17]：首先，将预测模型集成至监控平台，进行为期 1 个月的试运行与微调[3][20]；随后，在制冷系统自动控制回路中，以“建议值”形式引入 DDPG 智能体生成的控制设定点，由运维人员确认后执行[2][7]；最后，在系统运行稳定后，开启全自动闭环控制模式[6][12]。

4.2 运行效果分析

工程验证期间，数据中心实际运行的年平均 PUE 稳定在 1.32，与仿真结果偏差小于 4%，证明了仿真模型的高保真度与优化策略的有效性[5][6]。机房核心区域温度被控制在 22.5° C – 23.8° C 的稳定范围内，完全满足 IT 设备运行要求[15][17]。经财务测算，相较于优化前同期，该数据中心年节约用电量约 120 万千瓦时，折合电费约 82 万元人民币[9][13]。本次优化方案主要涉及软件开发与系统集成，硬件投入极少，总投资约 98 万元，其静态投资回收期约为 1.2 年，经济效益显著[11][18]。

5 结语

本文针对数据中心 PUE 优化中存在的协同性差、验证难、风险高等问题[1][5]，提出并验证了一套基于数字仿真模型的协同优化方法。通过构建高保真仿真环境，融合 Bi–LSTM 预测与 DDPG 强化学习算法[3][7][12]，实现了对 IT 负载、制冷系统的前瞻性协同调控[2][16]。仿真与长达 6 个月的工程实践均表明，该方法能显著降低 PUE（从 1.48 至 1.32）与制冷能耗（降低 22.4%），同时提升负载均衡度与温度稳定性[19][20]，且具有投资回收期短（1.2 年）、适应性强等特点[9][11]。本研究为数据中心的精细化节能管理提供了一条“仿真驱动、智能决策、安全落地”的可行路径[6][17]。未来工作可从以下方面展开：一是探索轻量化模型部署，在保证精度的同时降低算法对计算资源的需求，以适配边缘数据中心等场景[14][20]；二是将可再生能源出力、分时电价等因素纳入优化目标，实现综合运行成本最小化[18]；三是研究跨数据中心集群的联合优化调度，从系统级进一步提升能效水平，助力数字基础设施的绿色低碳转型[1][12]。

参考文献

[1] 中国信息通信研究院 . 数据中心白皮书（2024 年）[R]. 北京：中国信息通信研究院，2024.

[2] 张建，郭亮，王海峰 . 基于深度强化学习的云计算数据中心节能调度研究 [J]. 计算机学报，2023,46（5）：1020–1035.

[3] 李华，刘伟，陈明 . 基于 LSTM 和模型预测控制的机房空调系统节能优化 [J]. 制冷学报，2022, 43（4）：

85–93.

［4］王志强，肖鹏，赵宇 . 数据中心热环境模拟与气流组织优化［J］. 暖通空调，2021, 51（S2）: 156–161.

［5］国家市场监督管理总局，国家标准化管理委员会 .GB 40879–2021 数据中心能效限定值及能效等级［S］. 北京：中国标准出版社，2021.

［6］吴畏，孙磊，黄小华 . 基于数字孪生的数据中心基础设施管理系统设计与实现［J］. 电信科学，2023, 39（7）: 124–135.

［7］马超群，周洪波 . 强化学习在能源互联网优化控制中的应用综述［J］. 电力系统自动化，2020,44（21）: 179–191.

［8］宋晓，李俊，胡波 . 考虑多时间尺度的数据中心冷却系统优化运行策略［J］. 中国电机工程学报，2022,42（10）:3655–3665.

［9］阿里巴巴集团技术团队 . 数据中心绿色运维实践与思考［M］. 北京：电子工业出版社，2023.

［10］王振宇，刘芳，蔡毅 . 基于 CFD 仿真的数据中心冷通道封闭效果分析［J］. 工程热物理学报，2021,42（8）: 2011–2018.

［11］Uptime Institute. Global Data Center Survey 2023［R/OL］.（2023）.［2025–01–15］. https://uptimeinstitute.com/resources/asset/2023–data–center–industry–survey.

［12］Li Y, Wen Y, Tao D, et al. Transforming cooling optimization for green data center with deep reinforcement learning［J］. IEEE Transactions on Parallel and Distributed Systems, 2022, 33（4）: 1012–1025.

［13］Google. DeepMind AI Reduces Google Data Centre Cooling Bill by 40%［EB/OL］.（2018–08–20）［2025–01–20］. https://deepmind.google/discover/blog/deepmind–ai–reduces–google–data–centre–cooling–bill–by–40/.

［14］Wang C, Urgaonkar B, He Q, et al. Predictive and adaptive co–management of power and performance for virtualized data centers［J］. IEEE Transactions on Cloud Computing, 2021, 9（1）: 274–287.

［15］ASHRAE. Thermal Guidelines for Data Processing Environments［M］. 5th ed. Atlanta: ASHRAE, 2021.

［16］Zhang Q, Lin M, Yang L T, et al. A double deep Q–network learning approach for energy–efficient task scheduling in data centers［J］. Future Generation Computer Systems, 2020, 113: 226–239.

［17］The Green Grid. PUE ™ : A Comprehensive Examination of the Metric［R］. White Paper #49. Beaverton: The Green Grid, 2012.

［18］Gao J, Wang H, Shen H. Smartly handling renewable energy instability in green data centers with deep reinforcement learning［J］. IEEE Transactions on Sustainable Computing, 2023, 8（1）: 218–230.

［19］Patterson M K. The effect of data center temperature on energy efficiency［C］//Proceedings of the 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems. IEEE, 2008: 1167–1174.

［20］Dey S, Roy A, Das S K. A survey of energy–efficient data center management using machine learning［J］. ACM Computing Surveys, 2023, 56（2）: 1–38.

作者简介：胡涛（1984– ），男，汉族，陕西西安人，硕士，浙江政务服务专家库成员，主要研究方向为算力中心运维项目管理。