基于模型深度强化学习的数据中心主动地板控制

温建伟[1], 张　立[1†], 段彦夺[2], 李雷孝[2]

(1. 内蒙古自治区气象信息中心, 内蒙古 呼和浩特 010051;

2. 内蒙古自治区基于大数据的软件服务工程技术研究中心, 内蒙古 呼和浩特 010080)

摘要: 如何消除数据中心的局部热点是困扰数据中心行业的关键问题之一. 本文采用主动地板(AVT)来抑制局部机架热点现象, 并将数据中心AVT控制问题抽象为马尔可夫决策过程, 设计了基于深度强化学习的主动地板最优控制策略. 该策略基于模型深度强化学习方法, 克服了传统无模型深度强化学习方法采样效率低的缺陷. 大量仿真实验结果表明, 与经典无模型(PPO)方法相比, 所提出的方法可迅速收敛到最优控制策略, 并可以有效抑制机架热点现象.

关键词: 数据中心; 主动地板; 强化学习; 性能评价

引用格式: 温建伟, 张立, 段彦夺, 等. 基于模型深度强化学习的数据中心主动地板控制. 控制理论与应用, 2022, 39(6): 1051 – 1056

DOI: 10.7641/CTA.2021.10682

Model-based reinforcement learning for active ventilated tiles control in data centers

WEN Jian-wei[1], ZHANG Li[1†], DUAN Yan-duo[2], LI Lei-xiao[2]

(1. Inner Mongolia Meteorological Information Center, Hohhot Inner Mongolia 010051, China;

2. Inner Mongolia Autonomous Region Engineering & Technology Research Center of Big Data Based Software Service, Hohhot Inner Mongolia 010080, China)

Abstract: How to remove the hotspots in data centers is one of the key issues in the data center industry. This work focuses on designing active ventilation tiles (AVTs) to restrain hotspots in data centers. The AVT control problem is abstracted into a Markov decision process (MDP) problem, and an optimal control algorithm based on deep reinforcement learning (DRL) is proposed. The proposed approach adopts the model-based reinforcement learning (MBRL) paradigm and has better sample efficiency compared to traditional model-free approaches. Extensive simulation studies are conducted and numerical results show that our algorithm learns the optimal control policy faster than the classical model-free proximal policy optimization (PPO) algorithm and is effective in suppressing the local hotspots in data centers.

Key words: data center; active ventilation tile; reinforcement learning; performance evaluation

Citation: WEN Jianwei, ZHANG Li, DUAN Yanduo, et al. Model-based reinforcement learning for active ventilated tiles control in data centers. *Control Theory & Applications*, 2022, 39(6): 1051 – 1056

1　引言

近年来, 随着人工智能、5G通信技术、物联网的不断发展, 用户对存储、计算资源的需求不断增加. 数据中心作为给用户提供存储和计算服务的载体, 其数量在不断的上升. 然而, 由于数据中心空间受限, 数据中心不得不提高功率密度来降低数据中心运营成本. 但是随着数据中心功率密度提高, 散热已然成为数据中心的一个难题. 冷却问题成为当今数据中心的挑战.

目前关于数据中心冷却控制研究可以分为全局控制和局部控制. 在全局控制研究领域内, Lazic等人[1]提出了模型预测控制的方法应用在真实的数据中心中, 采用一种数据驱动的, 基于模型的强化学习方法用于调节大型数据中心内的温度和气流. 结果表明, 强化学习代理仅仅需要几个小时的探索就可以有

收稿日期: 2021−07−27; 录用日期: 2021−10−27.

†张立. E-mail: 845885886@qq.com; Tel.: +86 15049170010.

本文责任编委: 赵千川.

国家自然科学基金项目(61862048), 内蒙古自治区科技重大专项项目(2019ZD015, 2019ZD016), 内蒙古自治区关键技术攻关计划项目(2019GG273, 2020GG0094), 内蒙古自治区科技成果转化专项资金项目(2020CG0073, 2021CG0033)资助.

Supported by the National Natural Science Foundation of China (61862048), the Inner Mongolia Key Technological Development Program (2019ZD015, 2019ZD016), the Key Scientific and Technological Research Program of Inner Mongolia (2019GG273,2020GG0094) and the Inner Mongolia Special Program for Engineering Application of Scientific and Technical Researches (2020CG0073, 2021CG0033).

效、安全地调节数据中心的温度分布. 并且相比于比例–积分–微分 (proportional integral differential, PID)控制器的方式, 基于模型的强化学习方法大大提升了工作效率. Li等人[2] 提出了一种端到端的冷却控制算法 (cooling control algorithm, CCA), 该算法结合了Actor-Critic框架和深度确定性策略梯度算法(deep deterministic policy gradient, DDPG)的离线策略, 在真实的数据中心进行了评估. 经过验证CCA可以节省11%的冷却成本. Chi等人[3]提出了无模型强化学习(model free reinforcement learning, MFRL)算法MAD-3C(multi-agent drl-based data center cooperative control), 解决了数据中心能耗优化问题中状态空间和行为空间维数爆炸问题, 并且设计了演员–评论家的深度确定性策略梯度算法(actor critic deep deterministic policy gradient, AC-DDPG)多智能体合作框架, 用于改善IT系统和冷却系统之间的合作. 实验表明, 该方法能够在保证训练稳定性和提高资源利用率的同时, 通过协同优化有效降低数据中心的能耗. 以上相关研究均从数据中心级水阀, 空调风扇等全局因素进行数据中心全局制冷控制, 但是全局控制必然存在精细度不足问题, 例如, 在解决局部制冷问题时, 全局控制通常能效比较低, 无法达到预期效果[4]. 在局部控制研究领域内, Beitelmal等人[5]通过在数据中心冷通道安装AVT用于数据中心局部控制, 改善数据中心整体冷却效率. 为了减少数据中心能耗, Zhou等人[6] 设计了一种基于AVT的模型预测控制器, 以协调全局和局部冷却, 并最小化数据中心冷却功耗, 实验表明, 该方案可降低36%冷却功耗. 李永利等人[7]设计了AVT的热能效预测模型, 对基于AVT的控制提供了模型基础. Wan 等人[8–9]提出了基于强化学习的AVT控制问题, 分别使用了Q-Learning算法和深度Q网络 (deep Q network, DQN)算法, 通过在真实的数据中心验证, 结果表明AVT的应用不仅可以缓解数据中心中局部机架热点问题, 而且可以降低整个数据中心制冷能耗. 以上AVT控制算法使用的都是无模型的强化学习算法, 采样效率低, 算法收敛速度慢. 并且无论是Q-Learning算法还是DQN算法都只适用于离散行为空间问题, 而AVT控制更适合于连续行为空间问题, 所以上面研究都是将AVT的行为空间进行了离散化, 这不利于求解出AVT的最优控制策略.

为了解决上述问题, 本文提出了一种基于模型深度强化学习(model based reinforcement learning, MBRL)的方法解决数据中心机架级AVT控制问题. 主要贡献如下: 1) 提出了MBRL方法用于数据中心机架级AVT控制, 消除了局部机架热点问题. 并且对于AVT的控制采用其连续的行为空间, 实现了AVT的细粒度控制; 2) 对于MBRL方法, 本文在文献[1]的基础上对MBRL方法进行了改进. 将文献[1]中线性的环境模型改为非线性的神经网络结构模型, 使环境模型拟合更加精准. 另外引入了策略神经网络学习模型预测控制器给出的专家行为, 减少了整体决策的时间, 利于系统向前推进; 3) 通过在数据中心模型中进行仿真实验验证, 仿真结果证明MBRL算法相比于PPO算法收敛速度快, 并且MBRL算法相比于PPO算法节约了16%的功耗.

2 问题描述

数据中心普遍都是高架地板结构, 如图1所示.

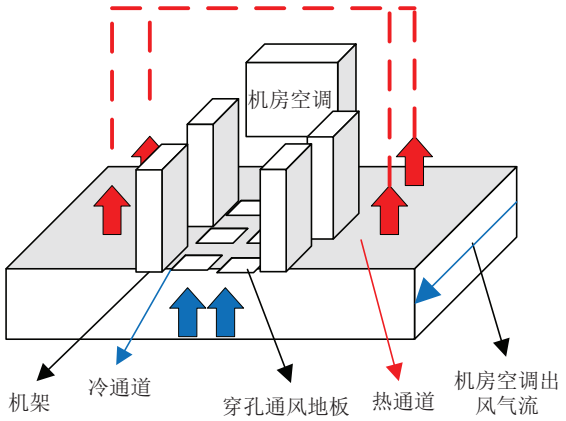

图 1 数据中心气流组织

Fig. 1 Data center air distribution

机架面对面形成冷通道, 背对背形成热通道. 机房空调(computer room air conditioner, CRAC)通过下送风的方式, 将产生的冷气输出到抬升地板的下气室. 冷气经过穿孔通风地板进入冷通道, 在服务器风扇的作用下穿过机架, 带走服务器运行产生的热量, 产生的热气流排进热通道. 热通道的热气汇合被朝上开口的空调吸入进行换热. 根据穿孔通风地板上是否安装风扇, 可分为AVT和被动地板(passive ventilated tiles, PVT)两类.

安装被动地板的数据中心容易出现局部热点现象, 即一些机架的一个或几个位置温度明显高于其他位置温度. 通常机架热点产生的主要原因有两个: 服务器运行负载过大. 冷气供应不足引起热气回流, 形成热点. 容易形成热点的位置包括一些冷通道末排机架、机架顶端、机房空调周围机架等[10]. 在这些特殊位置, 即使空调满荷运行, 热点现象依然存在[11]. 针对当前数据中心机架的局部热点现象, 在过热机架位置处的被动地板替换为主动地板, 采用基于模型深度强化学习的算法控制AVT上的风扇转速, 在不增加空调制冷能耗的前提下增强局部机架级冷气供应, 实现降低整体机架温度的目标.

3 马尔可夫决策过程建模

AVT的控制是序贯决策过程, 本文将AVT的控制问题建模为马尔可夫决策过程(Markov decision process, MDP). 其中包括状态空间行为空间和奖励.

状态空间: 采用机架进口温度分布作为系统状态. 通过直接在服务器机架的正面安装一组温度传感器来测量入口温度分布. 将温度传感器集合记为I, 传感器I在t时刻的读数记为$T_{t,i}$, 因此在t时刻的状态空间定义为向量$s_t = \{T_{t,i}\}, i \in I$.

行为空间: 通过控制主动地板上风扇转速实现对主动板的控制. 本文将风扇转速定义为一个连续的行为空间, 在t时刻风扇的转速为a_t, 其范围定义为$a_t \in A = [0, f_{\max}]$, 其中$f_{\max}$表示风扇最大转速.

奖励: 奖励表示在当前状态下采取行为所获得的收益. 本文的优化目标是消耗尽可能少的能源并且抑制局部热点, 据此定义奖励函数如下:

$$R_t = (1-\omega)R_{t,T} + \omega R_{t,E}. \tag{1}$$

奖励函数分为温度和功耗两部分组成, 温度部分定义如下:

$$R_{t,T} = -\frac{\sum_{i\in I}(T_{t,i} - T_{t,\text{threshold}})^2}{|I|}, \tag{2}$$

其中: $T_{t,\text{threshold}} = T_{\text{below}} + \tilde{T}$, T_{below}为每个时间步机房空调出风口测量得到的冷风送风温度, $\tilde{T}$为冷风在非密闭环境下传输过程中产生的温度升高, $|I|$为安装在机架上的传感器的数量. 当传感器温度$T_{t,i}$越接近参考温度$T_{t,\text{threshold}}$时, $R_{t,T}$越大, 即机架入风口温度分布越均匀. 根据风机定律[12], 风扇功耗部分的奖励$R_{t,E}$可定义为如下:

$$R_{t,E} = -(\frac{a}{A_{\max}})^3, \tag{3}$$

其中$a_t \in A$, $A_{\max}$为风扇转速最大值. 显然, 风扇转速越小, 风扇功耗越低, 风扇获得的功耗奖励$R_{t,E}$越大. 式(1)中ω为平衡$R_{t,E}$与$R_{t,T}$的权重. 由于R_t为负值, 使得R_t最大化即趋向于0是本文的优化目标.

4　基于模型深度强化学习算法设计

无模型深度强化学习算法具有广泛的问题适应性. 然而, 无模型深度强化学习算法受到样本采样效率非常低的局限性, 需要不断的与环境交互才能获得训练样本[13], 因此在实时控制问题中难于应用. 基于模型的强化学习算法通过学习环境模型可以直接与模型进行交互, 不需要直接和真实的环境进行交互, 因此样本采样效率更高.

4.1　算法设计

基于模型深度强化学习的AVT控制算法逻辑图如图2所示. 强化学习智能体, 通过策略网络选择行为a_t并执行到环境中, 环境产生下一个状态s_{t+1}, 并获得奖励r_t, 将智能体与环境交互产生的Experience(s_t, a_t, s_{t+1})存储到D_l中进行训练环境模型, 获取当前环境状态s_{t+1}通过环境模型与MPC控制器计算出在当前状态的最优行为a_t^*, 将s_t, a_t^*存储到D_e中进行训练策略神经网络.

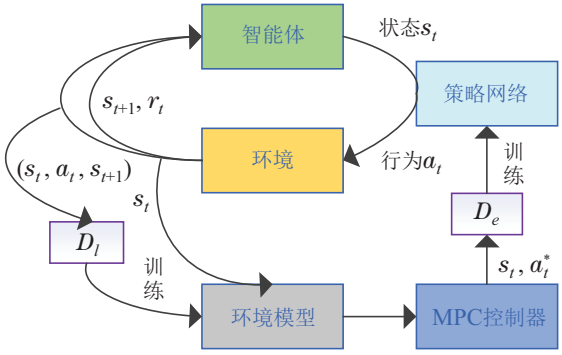

图 2　算法逻辑结构图

Fig. 2　Algorithm logical structure

基于模型强化学习的AVT控制算法如下所示:

1: 初始化E, L, 大小的经验回放池D_e, D_l; 随机初始化环境模型$\hat{f}_\theta(s,a)$参数和策略网络$\hat{f}_\phi(s)$;
2: 使用随机策略探索环境, 采集元组(s_t, a_t, s_{t+1})并存储到经验回放池D_l中, 预训练环境模型$\hat{f}_\theta(s, a)$;
3: **for** $t = 0; t \leqslant N; t++$ **do**
4: 　观察系统当前状态s_t;
5: 　使用ϵ-greedy策略选择行为

$$a_t = \begin{cases} \hat{f}_\phi(s), & 1-\epsilon\text{概率}, \\ \text{随机产生行为}, & \epsilon\text{概率}; \end{cases}$$

6: 　$\epsilon = \min(\epsilon - \Delta_\epsilon, \epsilon_{\min})$;
7: 　执行行为a_t;
8: 　观察当前系统的下一个状态, 计算即时奖励;
9: 　通过环境模型与MPC控制器求解出当前状态的最优行为序列$(A_t^{\mathrm{T}})^*$, 获取其中第1个行为$(a_t)^* \in (A_t^{\mathrm{T}})^*$;
10: 　将(s_t, a_t^*)存储到经验回放池D_e;
11: 　将(s_t, a_t, s_{t+1})存储到经验回放池D_l;
12: 　使用D_l训练$\hat{f}_\theta(s,a)$ K个回合;
13: 　使用D_e训练$\hat{f}_\phi(s)$ K个回合;

该算法第1步分别初始化经验回放池D_e, D_l, 用于存放经验样本, 分别用来训练策略网络$\hat{f}_\phi(s)$和环境模型$\hat{f}_\theta(s,a)$. $\hat{f}_\phi(s)$通过输入当前时刻的状态预测行为. $\hat{f}_\theta(s,a)$通过输入当前时刻系统的状态和行为, 预测出下一时刻系统的状态. 第2步使用随机策略对环境进行探索, 采集经验样本(s,a,s')存储到D_l中, 使用采集的经验样本预先训练$\hat{f}_\theta(s,a)$. 强化学习过程为第4–13步: 其中, 第4步观察系统当前状态. 第5–7步使用ϵ-greedy策略在当前状态选择行为并执行, 其中依概率ϵ在行为空间内随机产生行为, 依概率$1-\epsilon$使用策略网络生成行为, 算法起始阶段ϵ较大, 随着算法的不断迭代, ϵ逐渐递减, 算法逐渐倾向于选择策略网络

生成的行为. 第8步观察系统的下一个状态并计算执行行为之后得到的奖励. 第9步使用环境模型求解在当前系统状态下应该采取的最优行为(具体过程见第4.3节模型预测控制). 第10–13步存储样本数据并训练环境模型与策略网络. 重复迭代上述过程, 当算法收敛时策略网络将学习到近似最优解.

4.2 构建环境模型与策略网络

算法中的第5步需要利用策略网络$\hat{f}_\phi(s)$生成行为a_t. 本研究将$\hat{f}_\phi(s)$设计为一个深度全连接神经网络. $\hat{f}_\phi(s)$的输入是系统当前时刻的状态, 输出是行为变量. 为了约束输出的行为空间在$[0, f_{\max}]$, 本文使用Tanh激活函数将输出的行为规约到$[-1, 1]$的空间内. 在实施行为时, 再将$\hat{f}_\phi(s)$的输出反映射到行为空间A对应的范围. 算法中第9步需要利用环境模型$\hat{f}_\theta(s,a)$预测下一个时刻机架正面板的温度分布状态. 本研究将环境模型$\hat{f}_\theta(s,a)$设计为一个深度全连接神经网络, 其输入是当前时刻下的机架正面板的温度状态s_t和当前时刻下采取的风扇转速a_t. 该神经网络模型的目标是尽可能准确地预测未来状态, 因此使用未来预测状态与真实状态的均方差作为其损失函数, 公式定义如下:

$$\varepsilon(\theta)=\frac{1}{D}\sum_{(s_t,a_t,s_{t+1})\in D_l}\frac{1}{2}\|s_{t+1}-\hat{f}_\theta(s,a)\|^2, \quad (4)$$

算法的每次迭代过程中, 在经验回放池中选取小批量样本(MiniBatch)将$\hat{f}_\theta(s,a)$训练K次.

4.3 模型预测控制

算法的第9步是利用环境模型$\hat{f}_\theta(s,a)$进行行为序列规划求解奖励最大的行为序列, 定义如下:

$$\underset{a_t,a_{t+1},\cdots,a_{t+T-1}}{\arg\max}\sum_{t'=t}^{t+T-1}r(\hat{s}_{t'},a_{t'}), \quad (5)$$

本文采用Random Shooting[14]的方式. 随机生成M个维度为T的行为序列. 使用环境模型预测出每个行为序列的累积奖励, 然后将M个行为序列根据累计奖励进行排序, 得到累计奖励最高的行为序列, 如式(6):

$$(A_t^{\mathrm{T}})^*=\underset{A_t^{\mathrm{T}}}{\arg\max}\sum_{t'=t}^{t+T-1}r(\hat{s}_{t'},a_{t'}),\ a_{t'}\in A_t^{\mathrm{T}}, \quad (6)$$

将累积奖励最大的行为序列的第1个行为$(a_t)^*\in(A_t^{\mathrm{T}})^*$作为最优行为存入到$D_e$中. 而后, 可利用模仿学习训练策略网络, 即将$\hat{f}_\phi(s)$输出的行为与基于模型预测控制求解的行为的均方差作为损失函数训练策略网络, 如式(7):

$$\phi(\theta)=\frac{1}{|D|}\sum_{(s_t,a_t^*)\in D_e}\frac{1}{2}\|\hat{f}_\phi(s_t)-(a_t)^*\|^2, \quad (7)$$

算法的每次迭代过程中, 在经验回放池D_e中选取小批量样本(MiniBatch)将$\hat{f}_\phi(s)$重复训练K次.

5 仿真实验

5.1 实验环境

本文使用6SigmaDC进行CFD(computational fluid dynamics)仿真, 数据中心采用典型的冷暖通道方式排列. 数据中心内部环境具体设计如图3所示.

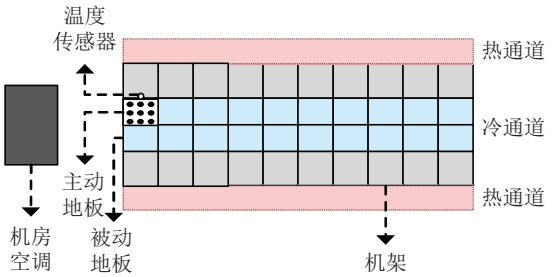

图 3 数据中心布局

Fig. 3 Data center layout

数据中心由两排机架组成, 其中每排放置了10个机架, 其中每个机架内放置了5个4 U大小的标准服务器, 每个服务器的功率是800 W, 每个机架的总体功耗是4 kW. 其中机房空调设置为送风温度控制, 温度设置为22°C, 风机转速设置为77%. 经过初步仿真发现由于接近机房空调的机架热气回流现象严重, 在该位置附近热点明显, 所以在第1个机架上部署主动地板进行实验.

主动地板是由被动地板和9个风扇组成, 其中每个风扇的风量取值范围为0到200的连续实数区间, 单位为立方英尺每分钟(cubic feet per minute, CFM). 在机架正面板上均匀放置6个传感器收集机架面板的温度. 在主动地板的下面放置一个传感器, 收集CRAC送风温度.

本文使用阿里云2018年发布的数据中心集群负载[15], 用于模拟真实的数据中心服务器负载变化. 该负载包括服务器CPU等资源利用率. 由于服务器CPU的利用率与热负载成线性关系[16], 本文将服务器CPU的利用率以线性方式映射为热负载. 取得前190 step的热负载变化曲线如图4所示.

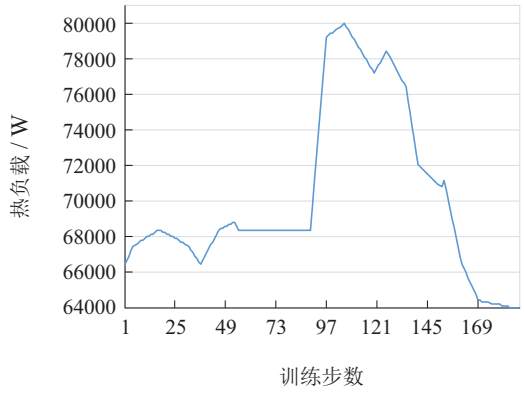

图 4 热负载

Fig. 4 Thermal load

5.2　算法参数设计

环境模型$\hat{f}_\theta(s,a)$设计为四层全连接神经网络. 其中3个隐藏层神经元个数分别为256, 128, 64. 激活函数使用ReLU. 优化器使用Adam, 学习率0.01. 策略网络$\hat{f}_\phi(s)$设计为四层全连接神经网络, 其中3个隐藏层神经元个数分别为128, 64, 32. 激活函数使用ReLU. 优化器使用Adam, 学习率为0.01. 算法其他主要参数如表1所示.

表 1　算法主要参数

Table 1　Main parameters of algorithm

参数	E	L	ϵ	Δ_ϵ	$\epsilon_{\min}$	ω
量值	64	128	1.0	0.04	0.1	0.7
参数	MiniBatch	γ	M	T	K	$\tilde{T}$
量值	32	0.9	200	4	32	2

5.3　算法有效性验证

通过在实验环境中运行本文提出的算法, 可以得出MBRL算法在抑制机架热点现象效果明显, 如图5红框标记机架所示, 其中上半部分左侧机架是使用被动地板时机架的温度分布, 可以明显看到机架中上部温度明显高于下部, 顶部温度达到33°C以上, 热点现象非常明显. 下半部分左侧机架是部署了主动地板之后机架的温度分布, 明显的是机架整体温度分布变得非常均匀, 整体温度分布在25°C, 部署了主动地板的机架热点现象消失. 所以通过部署主动地板可以改善机架的温度分布. 可以有效的抑制机架热点现象.

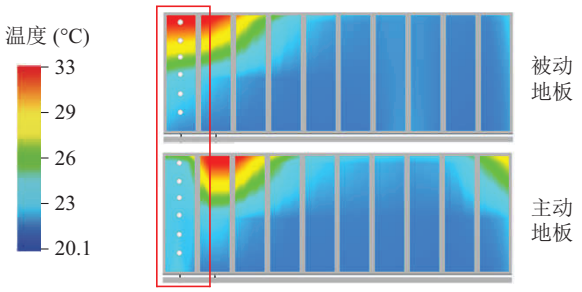

图 5　被动地板与主动地板的机架通风口温度

Fig. 5　The rack inlet temperature of pvt and avt

5.4　算法性能比较

为了验证本文算法的性能, 将本文算法与无模型深度强化学习(MFRL)算法PPO[14]进行了比较. PPO是一种策略梯度算法, 解决了策略梯度算法中步长难以确定的问题, 在目前无模型深度强化学习算法中性能较好. PPO算法是基于Actor-Critic结构, 其中Actor为选行为的策略神经网络, Critic是评价Actor选择的行为好坏的神经网络. 本文将基于PPO的AVT控制算法作为基线用于性能比较. 其中PPO算法中参数$\varepsilon=0.2$, Actor网络的学习率为0.01, Critic 网络的学习率为0.02. 将本文提出的MBRL算法与PPO算法进行了对比. 如图6即时奖励曲线所示.

随着算法的不断迭代, 奖励呈现上升的趋势. 从图中看到, 在50 step之前PPO算法震荡明显, 50 step之后算法开始收敛, 本文算法在15 step之后开始收敛, 所以本文算法前期的采样效率明显高于PPO. 在50～125 step之间两个算法都有较小的震荡, 在125 step之后两个算法的即时奖励几乎持平. 所以综合来看, 基于模型的强化学习算法整体比较平稳, 前期找到系统最优解的速度较快, 算法收敛速度更快.

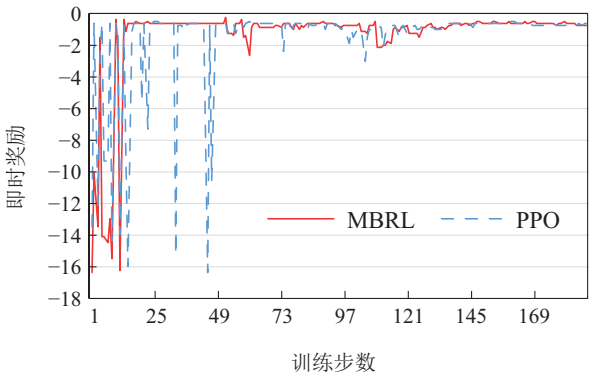

图 6　即时奖励

Fig. 6　The instant reward

如图7平均风扇转速所示: MBRL算法和PPO算法在算法运行期间的平均风扇转速分别为158 CFM和174 CFM, 因此MBRL算法运行期间的风扇功耗更低. 根据风机定律可知风扇转速可以间接体现出风扇功耗, 因此通过计算可得, 与PPO算法相比, MBRL算法节约了16%的风扇功耗.

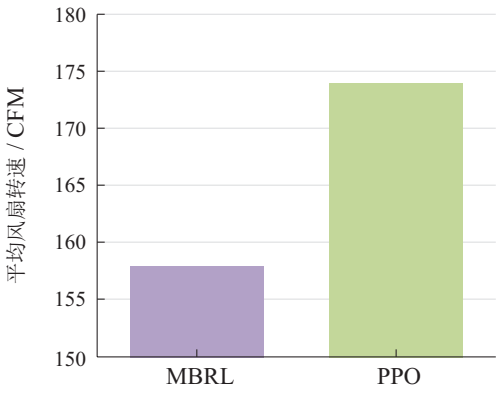

图 7　平均风扇转速

Fig. 7　Average fan speed

6　结语

本文研究了数据中心机架级AVT控制问题, 将AVT控制问题抽象为一个马尔可夫决策过程, 并设计了一种基于MBRL的AVT控制算法. 在数据中心模型中对MBRL算法进行了有效性验证, 并与无模型PPO算法进行了性能对比, 实验结果表明基于MBRL的

AVT控制算法的采样效率与学习速度明显优于PPO控制算法, 并且在抑制机架热点的同时, 降低了AVT功耗. 本文的研究内容, 为当今数据中心局部冷却与主动地板控制提供了参考, 对数据中心降低能耗, 抑制局部机架温度过高具有重要的实际意义.

参考文献:

[1] LAZIC N, BOUTILIER C, LU T, et al. Data center cooling using model-predictive control. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, Canada: ACM, 2018, 8: 3818 – 3827.

[2] LI Y, WEN Y, TAO D, et al. Transforming cooling optimization for green data center via deep reinforcement learning. *IEEE Transactions on Cybernetics*, 2019, 50(5): 2002 – 2013.

[3] CHI C, JI K, SONG P, et al. Cooperatively improving data center energy efficiency based on multi-agent deep reinforcement learning. *Energies*, 2021, 14(8): 1 – 32.

[4] WAN J, GUI X, KASAHARA S, et al. Air flow measurement and management for improving cooling and energy efficiency in raised-floor data centers: A survey. *IEEE Access*, 2018, 6: 48867 – 48901.

[5] BEITELMAL M H, WANG Z, FELIX C, et al. Local cooling control of data centers with adaptive vent tiles. *International Electronic Packaging Technical Conference and Exhibition*. San Francisco: InterPACK, 2009, 43604: 645 – 652.

[6] ZHOU R, WANG Z, BASH C E, et al. A holistic and optimal approach for data center cooling management. *Proceedings of the 2011 American Control Conference*. San Francisco: IEEE, 2011: 1346 – 1351.

[7] LI Yongli, NIU Hongxun, ZHOU Jie, et al. Research on model of active tiles in data centers based on machine learning. *Journal of Computer Simulation*, 2019, 36(12): 180 – 185.
(李永利, 牛弘勋, 周杰, 等. 基于机器学习的数据中心主动地板模型研究. 计算机仿真, 2019, 36(12): 180 – 185.)

[8] DUAN Y, WAN J, ZHOU J, et al. Reinforcement learning for rack-level cooling. *International Conference on Mobile Wireless Middleware, Operating Systems and Applications*. Hohhot: Springer, 2020: 167 – 173.

[9] WAN J, ZHOU J, GUI X. Intelligent rack-level cooling management in data centers with active ventilation tiles: A deep reinforcement learning approach. *IEEE Intelligent Systems*, 2021, 36(6): 42 – 52.

[10] DANG C, JIA L, LU Q. Investigation on thermal design of a rack with the pulsating heat pipe for cooling CPUs. *Applied Thermal Engineering*, 2017, 100(110): 390 – 398.

[11] FU L, WAN J, YANG J, et al. Dynamic thermal and it resource management strategies for data center energy minimization. *Journal of Cloud Computing*, 2017, 6(1): 1 – 16.

[12] YEO S, HOSSAIN M M, HUANG J C, et al. ATAC: Ambient temperature-aware capping for power efficient datacenters. *Proceedings of the ACM Symposium on Cloud Computing*. Seattle: ACM, 2014: 1 – 14.

[13] NAGABANDI A, KAHN G, FEARING R S, et al. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. *IEEE International Conference on Robotics and Automation (ICRA)*. Brisbane: IEEE, 2018: 7559 – 7566.

[14] BAKARAC P, KVASNICA M. Fast nonlinear model predictive control of a chemical reactor: a random shooting approach. *Acta Chimica Slovaca*, 2018, 11(2): 175 – 181.

[15] ALIBABA. Alibaba cluster workload traces: china: alibaba, 2018 (2021-9-4)[2021-7-13]. available: https://github.com/alibaba/cluster-data.

[16] WANG Jiye, ZHOU Biyu, ZHANG Fa, et al. Data center energy consumption models and energy efficient algorithms. *Journal of Computer Research and Development*, 2019, 56(8): 1587 – 1603.
(王继业, 周碧玉, 张法, 等. 数据中心能耗模型及能效算法综述. 计算机研究与发展, 2019, 56(8): 1587 – 1603.)

[17] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms, 2017(2017, 8-18)[2021-7-13]. available: https: //arxiv.org/abs/1707.06347v2, 2017.

作者简介:

温建伟 高级工程师, 目前研究方向为气象信息化建设, E-mail: shevawen@163.com;

张 立 高级工程师, 目前研究方向为信息化建设、新一代信息技术研究应用, E-mail: 845885886@qq.com;

段彦夺 硕士研究生, 目前研究方向为机器学习、数据中心冷却管理、绿色计算, E-mail: duanyanduo@163.com;

李雷孝 教授, 硕导, 目前研究方向为智能交通运输大数据、云计算与大数据分析, E-mail: llxhappy@126.com.