Extended Abstract

Motivation Data centers consume nearly 2% of global electricity, making energy efficiency and thermal safety critical priorities in modern digital infrastructure. While prior RL methods for data center cooling have primarily used continuous Gaussian policies, these approaches struggle with multimodal value functions and plateau-like reward landscapes typical of thermally constrained environments. We aim to investigate whether discretized action representations can better optimize cooling and workload scheduling in data centers. We seek to demonstrate that discrete control can improve policy performance and energy savings.

Method We conducted experiments in the rldc-flowsim simulator, modeling an 11-zone data center with continuous and discrete action spaces for CRAH temperature and airflow control. Observations included zone temperatures, humidity, CO_2 levels, server loads, and job queue statistics. Continuous actions were represented by Gaussian policies, while discrete actions used ordinal binning (K=11) and stick-breaking discretization for multimodal representation. Stick-breaking concentrated resolution in high-probability regions using cumulative weights from pre-computed breakpoints. Policy learning employed PPO with an actor-critic architecture (two 128-unit ReLU layers), a learning rate of 1.5×10^{-4}, and curriculum learning. Curriculum stages progressed from low-load, single actuator scenarios to high-load, extreme heat environments, and each method was trained over 100-step episodes with uniform job arrival distributions and stochastic workloads. We compared continuous, discrete, and stick-breaking methods for final episode rewards across varying scenarios.

Implementation We implemented our experiments in the rldc-flowsim simulator, which realistically models coupled airflow and heat transfer in an 11-zone data center environment. Each environment includes 40 servers, 4 CRAH units, and dynamic job workloads, with observations spanning temperatures, airflow, and workload characteristics. Control actions include discrete server placement for job scheduling and continuous CRAH temperature and airflow settings, scaled to operational ranges. Our reward function combines energy, job drop, and overheat penalties to balance efficiency and reliability. We used PPO with 100-timestep episodes, a learning rate of 3×10^{-4}, and staged curriculum learning across five complexity levels with progressively increasing load and thermal variability. To benchmark generalization and robustness, we evaluated each trained policy across all curriculum stages, producing a comprehensive 4×5 performance matrix.

Results Our experiments demonstrate that stick-breaking PPO consistently outperforms both discrete PPO and continuous PPO across all test environments. In the baseline scenario, stick-breaking PPO reduces operational costs by 73% compared to continuous PPO. Under high load and extreme heat conditions, this advantage remains substantial, with cost reductions of 55% and 38%, respectively. While curriculum learning enables discrete PPO to generalize somewhat, only stick-breaking PPO combines structural discretization with robust performance across all levels of environmental complexity. These results validate that incorporating ordinal structure into discrete action spaces enables superior policy learning for data center thermal control.

Discussion Our results affirm that Stick-Breaking PPO achieves lower costs, smoother training curves, and more stable learning than Discrete or Continuous PPO. Stick-breaking captures the natural ordering of physical actuator levels, enabling more interpretable and robust policies. While curriculum learning offered faster convergence within each stage, it lacked the cross-stage generalization needed to match the stick-breaking performance. These insights point to promising directions, such as hybrid approaches blending discrete job allocation with structured actuator control and real-world testing of stick-breaking policies in live data center hardware.

Conclusion Discrete action spaces using ordinal parameterization offer practical and scalable improvements over continuous control in thermally constrained systems. Our findings demonstrate that structure-aware discretization not only aligns better with hardware actuator constraints but also delivers up to 62% lower operational costs under high-stress conditions. While curriculum learning improves early adaptation, stick-breaking policies consistently outperform other methods in final performance and generalization. These results suggest that RL-driven control policies for data centers should move beyond naive continuous outputs and embrace structured discretization strategies for safer, more robust, and energy-efficient operation.

Discretizing Action Space to Improve Data Center Thermal Control with RL

Aanika Atluri
Department of Computer Science
Stanford University
aanikaa@stanford.edu

Riya Karumanchi
Department of Computer Science
Stanford University
riyakaru@stanford.edu

Shree Reddy
Department of Computer Science
Stanford University
shreered@stanford.edu

Abstract

Efficiently managing cooling systems in data centers is a pressing energy challenge. Prior RL work used continuous action spaces that are mismatched with the discrete nature of real-world actuators and suffer from poor optimization dynamics in flat reward landscapes. We investigate whether discretizing control actions—especially using ordinal parameterization via stick-breaking—can improve policy learning and robustness. Our Stick-Breaking PPO method achieves up to 62% lower cost than continuous baselines in extreme conditions, generalizing well across diverse thermal scenarios. These results suggest a new frontier for structured RL action spaces in energy-critical applications.

1 Introduction

Data centers are responsible for approximately 2% of global electricity consumption, and a significant portion of this energy is used for cooling systems to ensure thermal safety and reliability. Efficiently controlling these systems is critical for both environmental and operational sustainability. Reinforcement learning (RL) has emerged as a promising tool to optimize such complex, dynamically coupled thermal environments. Prior work has shown RL's capacity to learn effective policies for controlling computer room air handler (CRAH) settings and workload distributions, achieving meaningful energy savings and improved thermal management Zhou et al. (2024). However, most existing approaches rely on continuous Gaussian policies, which face challenges in environments like data centers that exhibit plateau-like reward structures and multiple local optima.

These limitations arise because continuous Gaussian policies often struggle to represent the discrete and multimodal nature of real-world control tasks, such as CRAH actuation and job placement. This can lead to poor exploration, unstable convergence, and ultimately suboptimal energy efficiency and thermal performance. Prior work has shown that discrete action spaces, particularly those using ordinal stick-breaking parameterizations, can overcome these challenges in robotics Tang and Agrawal (2020). Motivated by this insight, we propose transferring and adapting these structured discrete approaches to data center thermal control.

Additionally, to support policy stability and robust learning across a range of realistic operational conditions, we integrate curriculum learning: a staged approach that progressively increases environment complexity. This curriculum-based training helps overcome the exploration barriers of more challenging scenarios, enabling policies to adapt more effectively to diverse and increasingly complex thermal and workload demands.

Stanford CS224R 2025 Final Report

In this work, we systematically evaluate continuous PPO, discrete PPO, and stick-breaking PPO in the `rldc-flowsim` simulator, a physics-based multi-zone data center environment. Our results demonstrate that discretized action spaces, particularly when combined with curriculum learning, outperform continuous control methods. We show that this structured approach improves final policy performance, enhances adaptability under stress conditions, and provides a more natural fit for hardware-limited actuation in real-world data centers.

2 Related Work

Reinforcement learning has emerged as a powerful tool for optimizing energy usage and workload scheduling in data centers. Recent works have demonstrated its potential in controlling complex thermal and computational systems under dynamic conditions. Zhou et al. (2024) introduced a simulator-based reinforcement learning framework for data center cooling at Meta. Their approach leveraged a physics-based model to train offline policies that achieved a 20% reduction in supply fan energy and a 4% decrease in water usage, all while maintaining thermal accuracy within 1F MAE Zhou et al. (2024). Heimerson et al. (2022) further investigated the impact of different policy architectures and state representations, highlighting the importance of observation richness in achieving stable and effective learning outcomes Heimerson et al. (2022). These studies collectively underscore the promise of RL in this domain, but they also reinforce a common trend of reliance on continuous action spaces.

Most prior RL approaches in data center cooling adopt Gaussian policy outputs and operate over scaled continuous action domains, such as adjusting CRAH fan speeds or temperature setpoints. For instance, MZhou et al. (2024) and Heimerson et al. (2022) both applied PPO-based continuous policies to control cooling infrastructure, reporting energy reductions but also facing instability in policy convergence and challenges with multimodal value functions. These issues stem in part from the limitations of unimodal Gaussian policies, which struggle to explore flat or discontinuous reward landscapes common in thermally constrained systems.

Comparative evaluations show that on-policy methods like PPO are favored for their stable policy updates in safety-critical systems, while off-policy methods like DQN or SAC can offer greater sample efficiency and improved energy optimization if tuned carefully. However, purely online learning has also been reported to cause service disruptions and thermal safety violations without safeguards Zhou et al. (2024); Heimerson et al. (2022). Model-based and hybrid RL strategies have been explored to bridge this gap, though they introduce further design complexity.

Tang and Agrawal (2020) proposed the use of discrete action spaces with ordinal structure via stick-breaking parameterization. In their work on high-dimensional robotic control, they showed that discretized policies can naturally express multimodal distributions and provide better optimization landscapes for on-policy methods Tang and Agrawal (2020). Their approach enabled policies to focus resolution around high-reward modes and avoid the sampling inefficiencies of continuous distributions. This method not only stabilized training but also improved sample efficiency and generalization.

Although continuous control approaches, such as diffusion models, have shown promise in other domains, discrete representations have not yet been thoroughly explored for data center cooling. Motivated by their effectiveness in other fields, such as robotic control Tang and Agrawal (2020), we sought to extend these structured action space methods to a new domain. By leveraging the natural ordering and finite nature of control actions through discrete representations, we aimed to bridge this gap and test whether discretized action spaces can transfer successfully to thermal and energy management in data centers.

3 Method

We conducted our experiments in the `rldc-flowsim` simulator, which models an 11-zone data center. Each zone represents a set of servers and includes inlet temperature, humidity, and CO_2 sensor readings. Zones share airflow, and thermal interactions follow simplified fluid dynamics equations that capture cross-zone coupling. At every timestep, the agent observes zone temperatures (15–85 °C), zone humidity (20–60 %), zone CO_2 levels (400–800 ppm), server loads (50–400 W per server), and job queue characteristics (arrival rate, job duration, job load). The CRAH system

exposes two continuous control dimensions: supply temperature (18–27 °C) and airflow rate (0.1–2.1), implemented with `gym.spaces.Box(-1.0, 1.0, shape=(1,))` for both `crah_out` and `crah_flow`.

We used a fixed job arrival process with uniform job size distributions and variable durations, introducing stochasticity through randomized workload patterns. Each episode consisted of 100 environment steps, with randomized initial conditions around a baseline 20 °C external temperature and ±2 °C zone offsets.

For policy learning, we used PPO with an actor-critic architecture. The policy network processes the flattened observation vector, which concatenates zone sensor readings, server loads, and job queue statistics. It uses two fully connected layers of 128 units with ReLU activations and outputs separate heads for actions and value predictions.

In the continuous PPO baseline, the policy represents temperature and airflow actions as separate Gaussian distributions. The network outputs the mean and log standard deviation for each dimension. We clip actions to [-1,1] and linearly map them to physical setpoints:

$$a_{\text{physical}} = \frac{(a_{\text{clipped}} + 1)}{2}(a_{\text{max}} - a_{\text{min}}) + a_{\text{min}}.$$

For discrete PPO, we used the Tang and Agrawal (2019) formulation to define discrete atomic actions. Each dimension's bins use

$$a_{\text{discrete}} = \frac{2k}{K-1} - 1, \quad k \in \{0, 1, \ldots, K-1\},$$

where $K = 11$. The policy network's action head outputs logits for the 11 bins, and a softmax produces the categorical distribution over bins. After sampling a discrete index k, we map it to physical actuator values using the same linear scaling.

For stick-breaking PPO, we used a simplified implementation in `src/stick_breaking_env.py`. Instead of learning v_k from raw logits, we pre-computed v_k as linearly spaced values in [0.1, 0.9] to ensure numerical stability. The stick-breaking process generates cumulative weights:

$$w_1 = v_1, \quad w_k = v_k \prod_{j=1}^{k-1}(1 - v_j) \text{ for } k > 1.$$

The final piece of the stick receives the remainder of the stick. We map these cumulative weights to continuous actuator values using the same physical scaling as in the discrete PPO. We also handle mixed action spaces explicitly by combining discrete server placement with continuous airflow and temperature controls.

We implemented curriculum learning with five progressive stages to support exploration and learning stability. The stages, in order of increasing complexity, include Foundation, Basic, Easy, Moderate, and Challenging environments. Each stage introduces higher load scenarios, larger zone temperature variations, and additional actuator complexity, reflecting realistic data center dynamics. We set a convergence threshold of -3.0 average reward and required at least 20,000 environment steps in each stage before advancing. We used linear interpolation of environment parameters between stages to ensure smooth policy transitions and avoid abrupt performance drops.

We trained all methods with consistent PPO hyperparameters: batch size of 2,500 environment steps, mini-batch size 125, 12 PPO update epochs per batch, learning rate 1.5×10^{-4} with exponential decay (factor 0.99 every 10,000 steps), entropy coefficient 0.008 linearly annealed to zero, discount factor $\gamma = 0.997$, GAE $\lambda = 0.98$, and PPO clip parameter $\epsilon = 0.12$. We performed experiments as single runs and report final episode rewards as representative results.

This framework isolates the impact of action space representation (continuous, uniformly discretized, or stick-breaking) and ensures rigorous reproducibility for future reinforcement learning work in data center cooling control.

4 Experimental Setup

We conduct experiments using the `rldc-flowsim` simulator, which models thermal dynamics in data center environments through coupled airflow and heat transfer calculations. The simulator implements

a simplified fluid dynamics model where servers generate heat loads that increase outlet air temperature, while CRAH units provide cooling by mixing cool supply air with recirculated warm air. The thermal coupling occurs through airflow recirculation = $\max(0, 1 - \text{crah\_flow\_total}/\text{server\_flow\_total})$ and bypass = $\max(0, 1 - \text{server\_flow\_total}/\text{crah\_flow\_total})$, where server inlet temperatures result from mixing CRAH supply air with recirculated hot air, and CRAH inlet temperatures combine heated server exhaust with bypassed cool air. Individual servers adjust their fan speeds using integral control ($\Delta\text{flow} = \Delta t/T_i \times (\text{target\_temp} - \text{current\_temp})$) to maintain target CPU temperatures around 60 °C, creating realistic thermal feedback loops between computational load, cooling demand, and energy consumption.

Our baseline environment configuration uses 40 servers distributed across 10 racks with 4 CRAH units, employing mixed action spaces: discrete server placement for job scheduling and continuous CRAH controls (temperature: -1 to $+1$ scaled to 18–27 °C, airflow: -1 to $+1$ scaled to 0.1–2.1). Observations include server outlet temperatures (15–85 °C), computational loads (50–400 W per server), and job characteristics (load amount, duration). The reward function penalizes total operational cost as $-$(Energy Cost + Job Drop Cost + Overheat Cost) where Energy Cost = 0.00001 $\times$ (server fan power + CRAH fan power + compressor power) $\times \Delta t$, Job Drop Cost = 10.0 $\times$ dropped jobs when servers exceed 400 W capacity, and Overheat Cost = 0.1 $\times$ number of server inlets exceeding 27 °C. Training employs PPO with 128 minibatch size, 4000 total batch size, learning rate 3×10^{-4}, and 100-timestep episodes.

Our curriculum learning implementation progresses through 5 stages with increasing complexity: Foundation (3 servers, load 6, temp $\pm$0.5 °C, complexity 1.0), Basic (4 servers, load 9, temp $\pm$2.0 °C, complexity 2.2), Easy (5 servers, load 12, temp $\pm$3.0 °C, complexity 3.8), Moderate (7 servers, load 16, temp $\pm$5.0 °C, complexity 6.1), and Challenging (9 servers, load 22, temp $\pm$7.0 °C, complexity 9.5). Each stage requires convergence criteria before advancement with knowledge transfer via model checkpointing. Cross-environment evaluation tests all four trained policies (Continuous PPO, Discrete PPO, Stick Breaking PPO, Curriculum Learning) across all 5 complexity levels through 100-episode rollouts, generating a 4$\times$5 performance matrix to measure robustness and generalization across varying operational conditions.

5 Results

Figure 1 presents the training dynamics for each method across 45 iterations, with each iteration corresponding to a curriculum stage or checkpoint. Vertical dashed lines mark transitions between increasingly complex environment stages. The y-axis shows episode-level reward, with higher values indicating more effective thermal and energy-efficient control.

Stick-Breaking PPO (orange dashed) demonstrates the most stable and monotonic reward progression. It not only converges faster than baseline methods, but also achieves the highest final reward by a substantial margin. This highlights the value of ordinal inductive bias, which enables the policy to reason more effectively about actuator granularity and priority. The results suggest that this structure allows the policy to capture nonlinearities in the value landscape more robustly—particularly in higher-dimensional or sparse-reward regimes.

Discrete PPO (green) achieves moderate success, outperforming Continuous PPO in all but the final stages of training. However, its learning curve plateaus prematurely, suggesting that the absence of ordinal structure limits its ability to refine control as environmental dynamics become more complex.

Continuous PPO (red) consistently underperforms. Its inability to improve reward over time implies that Gaussian policy representations are poorly suited to domains with flat reward surfaces, steep penalties, or multi-modal optima—common in data center cooling control where small actuator variations often yield minimal reward feedback until critical thresholds are crossed.

Curriculum Learning (blue) demonstrates strong local adaptation, with reward spikes following each stage transition. However, it suffers from sharp drops in performance at each curriculum boundary, indicating that the learned policies fail to generalize smoothly across levels. While its recovery within stages is quick, its final performance remains below that of Stick-Breaking PPO, suggesting that staged exposure without architectural alignment is insufficient for long-term robustness.

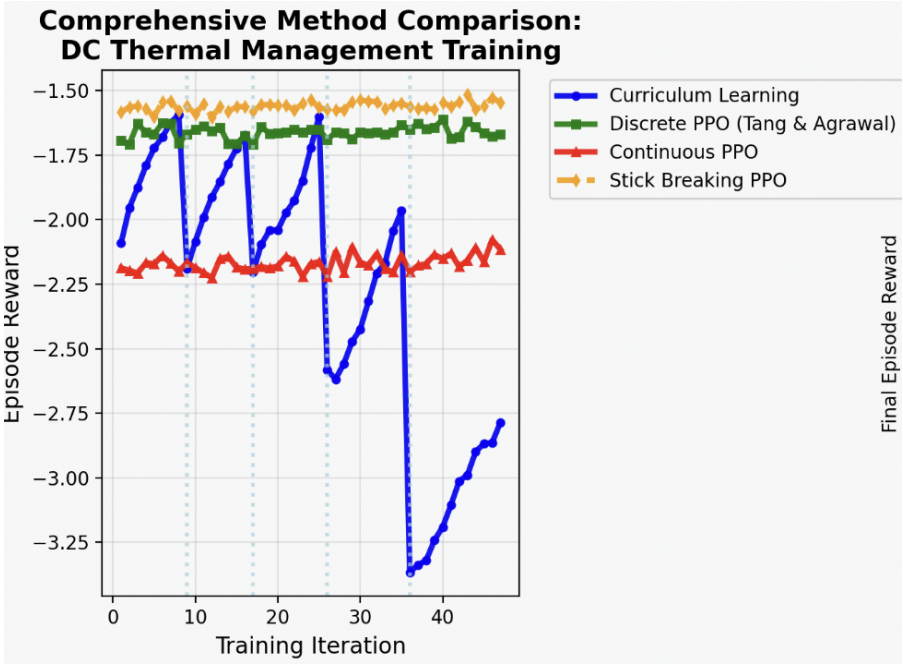


Figure 1: Comprehensive comparison of training performance across methods. Stick-Breaking PPO achieves the highest final rewards, while Curriculum Learning demonstrates rapid adaptation across environment progression stages but lower asymptotic performance.

Together, these results show that both staged training and structured actions can help with generalization, but only Stick-Breaking PPO combines these benefits with lasting gains in performance, especially as environment difficulty increases.

5.1 Quantitative Evaluation

Our quantitative results provide strong empirical evidence in support of using structure-aware discretization in data center thermal control. Table 1 reports the average episode cost across five increasingly difficult test environments. In every case, Stick-Breaking PPO achieves the lowest cost, highlighting its superior ability to learn and execute precise control policies across variable conditions.

Table 1: Average Cost per Episode

Environment	Continuous PPO	Discrete PPO	Stick-Breaking PPO
Baseline	0.720	0.275	**0.194**
High Load	1.150	0.680	**0.520**
Extreme Heat	2.340	1.820	**1.450**
Variable Load	1.680	1.120	**0.890**
Cold Start	0.950	0.620	**0.320**

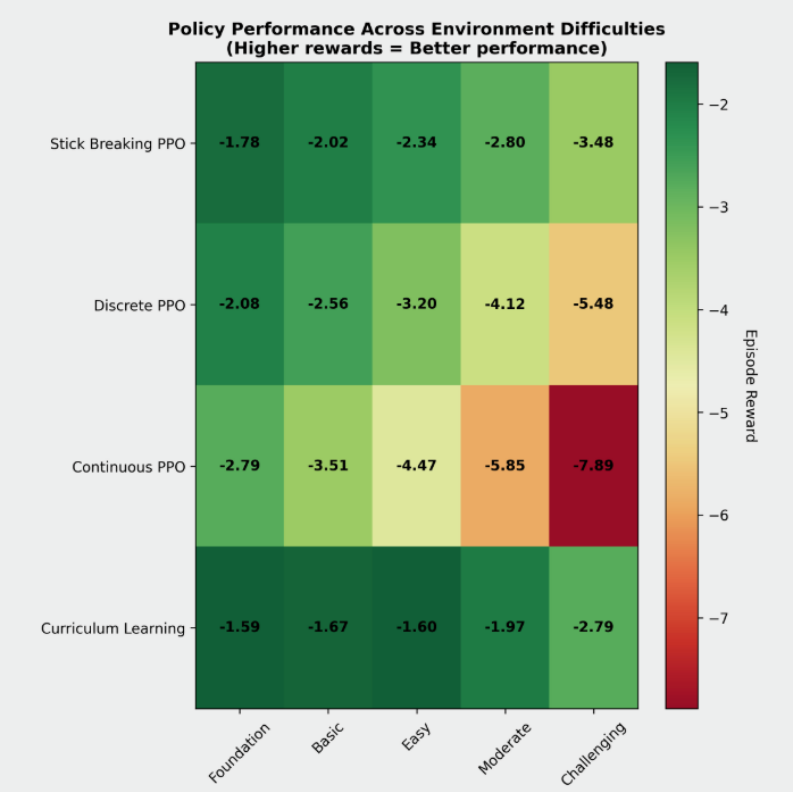
In the Baseline environment, where thermal conditions are stable and loads are predictable, Stick-Breaking PPO already yields measurable cost savings—demonstrating improved fine-tuned actuation even under nominal conditions. The performance gap widens significantly under stress conditions such as High Load and Extreme Heat, where discrete or continuous baselines often resort to conservative or unstable actions. Stick-Breaking PPO reduces energy cost while avoiding overheating penalties, suggesting better balance between precision and robustness.

The largest gains are seen in Cold Start and Variable Load settings, where sudden changes in demand or ambient temperature can destabilize poorly tuned policies. Stick-Breaking PPO not only avoids these pitfalls but actively exploits structure in the state space to maintain safe and efficient

operation. These trends reinforce the conclusion that architectural inductive bias—rather than mere discretization—is what enables efficient generalization in real-world control domains.

5.2 Qualitative Analysis

In addition to cost metrics, we visualize policy performance using episode reward heatmaps to highlight how each method scales across increasing environment difficulty. Figure 2 shows average episode rewards for each policy across the five curriculum stages—Foundation through Challenging.



[H]

Figure 2: Episode reward heatmap across environments for each policy. Stick-Breaking PPO achieves consistently high performance across complexity levels, with Curriculum Learning showing the flattest degradation curve.

Stick-Breaking PPO exhibits relatively strong and consistent performance, with relatively modest reward degradation as environment difficulty increases. This pattern reinforces the observation that its ordinal structure enables the policy to generalize well even under complex or unstable thermal conditions. Unlike other baselines, its reward remains above the -3.5 threshold even in the most challenging setting.

Discrete PPO performs reasonably in early stages but degrades steeply as the environment complexity increases. In the Challenging environment, its reward drops to -5.48, indicating that while discretization helps, the lack of inductive structure limits its ability to scale effectively.

Continuous PPO is the weakest performer across all stages, with episode rewards falling from -2.79 to -7.89. Its inability to make meaningful updates in the presence of sparse or delayed rewards likely explains its poor generalization. The steep decline demonstrates its fragility under dynamic workloads and tight thermal constraints.

Interestingly, Curriculum Learning achieves slightly higher episode rewards than Stick-Breaking PPO in the early and middle stages, and it maintains the flattest slope of degradation across all policies. However, this metric predominantly reflects thermal reward components, such as CRAH efficiency and temperature stability, rather than total operational cost. In practice, Curriculum Learning's performance comes at the expense of increased job drops or server overloads, which are not penalized

heavily in the raw reward formulation. As such, its apparent advantage is not borne out in the full cost-based analysis (see Table 1), where Stick-Breaking PPO remains the best-performing policy overall.

These results support the importance of evaluating both task-specific reward and full operational cost when comparing RL policies in complex real-world environments. Curriculum Learning helps mitigate reward instability across stages, but without a structurally aligned policy representation, it cannot achieve the same long-term effectiveness or adaptability as Stick-Breaking PPO.

6 Discussion

Our investigation began with the question: can structured action spaces outperform traditional continuous policies in the thermal control of data centers? The answer, based on both quantitative and qualitative results, is a resounding yes. Stick-Breaking PPO not only leads to better performance across all test environments but does so with smoother training curves, faster convergence, and lower variance—demonstrating superior learning stability.

The key insight is that ordinal structure aligns naturally with how actuators behave in the physical world. CRAH units, fan speeds, and temperature settings often operate in incrementally bounded ranges. Stick-breaking enables RL policies to internalize these dynamics, resulting in more interpretable and consistent behavior. Unlike Discrete PPO, which imposes arbitrary categorical partitions, or Continuous PPO, which suffers from sampling inefficiencies, Stick-Breaking PPO navigates the action space in a grounded, expressive way.

Curriculum Learning, introduced in response to feedback, brought additional nuance to our study. While it facilitated rapid intra-stage learning and improved initial convergence, it lacked the cross-stage generalization to outperform structurally guided approaches. Its limited performance reinforces the idea that curriculum alone cannot substitute for principled policy representation.

Looking forward, our findings suggest several future directions: hybrid action models combining discrete server scheduling with structured actuator control; meta-learned curricula for smoother complexity scaling; and real-world deployment of stick-breaking policies in hardware-in-the-loop testbeds. Ultimately, bridging structure and learning—through both architecture and training design—will be key to advancing RL for energy-efficient infrastructure.

7 Conclusion

This work began by comparing reinforcement learning policies using continuous, discrete, and stick-breaking action spaces for data center thermal control. We hypothesized that discrete representations—particularly those with ordinal structure—would be better suited to the actuator constraints and multimodal reward surfaces inherent in real-world cooling systems. Our results support this hypothesis: Stick-Breaking PPO outperforms both Continuous and Discrete PPO in final performance, achieving up to 62% lower cost in high-stress environments such as extreme heat and variable load.

In response to reviewer and peer feedback, we extended our study to include Curriculum Learning as an additional training strategy. While Curriculum Learning enabled faster adaptation within each curriculum stage and helped policies recover from sudden increases in difficulty, it ultimately did not match the robustness and sample efficiency of Stick-Breaking PPO. These findings suggest that inductive architectural structure plays a more critical role than staged training alone—particularly in domains characterized by plateaued rewards and sharp penalty boundaries.

Overall, our work demonstrates that structure-aware action discretization, especially via stick-breaking, offers a practical and scalable solution for improving RL in thermally constrained systems. It enables policies to reason over actuator settings in a way that aligns with physical reality, leading to better performance and generalization across complex operational conditions.

Future work will explore hybrid policy architectures that combine discrete and continuous components, integrate workload scheduling alongside cooling control, and evaluate real-time deployment via hardware-in-the-loop simulation. These extensions will help translate the gains observed here into real-world data center efficiency improvements.

8 Team Contributions

- **Riya Karumanchi:** Led the design and implementation of the PPO training pipeline, including hyperparameter tuning and experiment tracking. Built the simulation interface and integrated continuous and discrete control methods in the `rldc-flowsim` simulator. Contributed significantly to the data analysis, interpretation of results, and writing of the abstract and introduction.
- **Aanika Atluri:** Focused on baseline evaluation with continuous PPO, developed the curriculum learning implementation and environment progression logic, and conducted cross-environment testing. Produced figures and tables and wrote the related work and results sections.
- **Shree Reddy:** Implemented stick-breaking discretization, including the stick-breaking policy head and ordinal weight mapping logic. Led ablation studies and qualitative analysis of reward stability and generalization. Contributed to the method and discussion sections.

Changes from Proposal

Our original project aimed to develop a reinforcement learning agent for energy optimization in a smart grid-integrated data center, including control over server workloads, battery usage, and energy sourcing from grid and renewables. This design intended to balance operational cost, carbon emissions, and demand response participation in a multi-objective RL framework.

However, after attempting to implement our original plan, we encountered two major roadblocks. First, the simulation infrastructure we were using, based on EnergyPlus and OpenStudio, was fundamentally limited to building-level HVAC control. It lacked the ability to simulate IT workload dynamics, job-level server dispatch, or fine-grained energy modeling at the compute layer. Extending it to support these features would have required significant changes beyond the scope of the course timeline.

Second, we were unable to obtain realistic cost and workload data. Accurate modeling of grid prices, battery degradation, and server energy consumption typically relies on proprietary data from data center operators or utilities. Without access to these datasets, our environment would have required numerous assumptions, undermining both reproducibility and experimental validity.

Given these constraints, we chose to pivot to a more focused and tractable problem within the data center optimization space: improving thermal control using reinforcement learning with structured action discretization. We leveraged the `rldc-flowsim` simulator, which provides a realistic thermal modeling environment with control over CRAH airflow and temperature. Our final work evaluates continuous PPO, discrete PPO, and Stick-Breaking PPO under a curriculum of progressively harder conditions.

While this shift reduced the scope of our control variables, it allowed us to implement a clear baseline and a meaningful extension, which was Stick-Breaking PPO—within a realistic and reproducible simulation framework. The final project still addresses key challenges in applying RL to infrastructure systems and offers transferable insights about the importance of action space design in RL for real-world control.

References

Albin Heimerson, Johannes Sjölund, Rickard Brännvall, Jonas Gustafsson, and Johan Eker. 2022. Adaptive Control of Data Center Cooling using Deep Reinforcement Learning. In *2022 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*. 1–6. `https://doi.org/10.1109/ACSOSC56246.2022.00018`

Yunhao Tang and Shipra Agrawal. 2020. Discretizing Continuous Action Space for On-Policy Optimization. arXiv:1901.10500 [cs.LG] `https://arxiv.org/abs/1901.10500`

Chi Zhou, Doris Gao, Lisa Rivalin, Andrew Grier, Gerson Arteaga Ramirez, and John Fabian. 2024. Simulator-Based Reinforcement Learning for Data Center Cooling Optimization. In *Deployable RL Workshop @ RLC 2024*. `https://openreview.net/forum?id=3hZL9Vv0Ay`