Grid-Interactive Building Control Via Reinforcement Learning

Xiangyu Zhang
Computational Science Center
National Renewable Energy Laboratory (NREL)
Golden, CO, USA

The 13th International Conference on Applied Energy (ICAE 2021)
Nov. 29th, 2021

Motivation

- Abundant untapped demand side resources for grid services from the building sector.
- Effective and affordable smart building controllers suitable for mass deployment are yet to come.
- Among three major building appliances, the heating, ventilation and air-conditioning (HVAC) system poses more complexity for control.
- Mainstream grid-interactive building HVAC controllers are based on either direct load control (DLC, see SDG&E example [1]) or model predictive control (MPC) [2].
  - DLC: +: Easy to implement, affordable
    -: Does not directly consider building thermal condition.
  - MPC: +: Optimality for both building and grid control objective.
    -: Affordability: high costs for hardware (on-demand computation), software, modeling (accurate but simple building model) and maintenance.

Overview

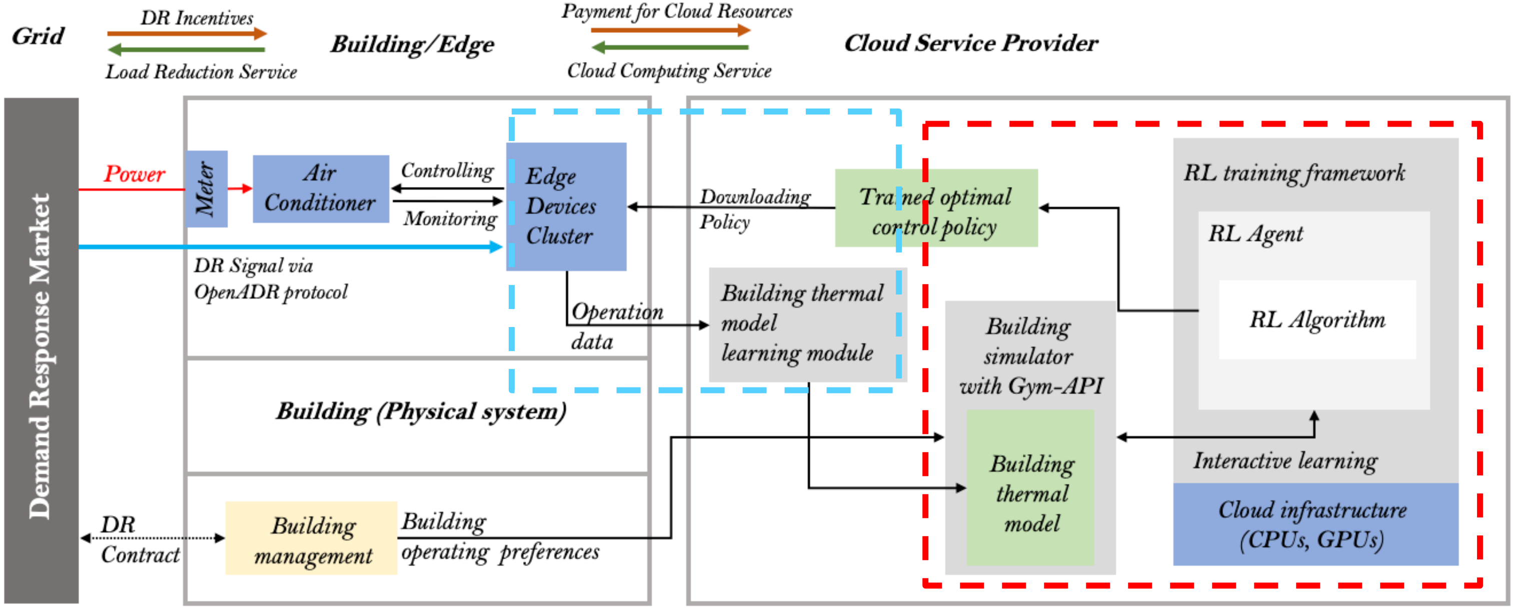

Fig. 1. Envisioned edge-cloud integrated solution for smart building grid-interactive control. System identification, controller training and real-time execution are automated based on such paradigm. [3]

Problem Formulation

- We investigate using deep reinforcement learning to solve a multi-zone grid-interactive building HVAC control problem with a continuous action space, the most complex single building control problem studied in RL literature.

Mathematical formulation for the optimal control problem:

$$\underset{\mathbf{a}_t \in \mathcal{A}, \forall t}{\text{minimize}} \quad \sum_{t \in \mathcal{T}} \mathbf{w}_t^{\top} \left[\kappa_1 \sum_{i=1}^{N} \mathcal{D}(T_t^i), \kappa_2 \mathcal{E}_t, \kappa_3 \mathcal{V}_t\right]$$

$$\text{subject to} \quad \mathbf{T}_{t+1} = \mathcal{F}(\mathbf{T}_t, \mathbf{a}_t, \varrho_t) \quad (\forall t \in \mathcal{T})$$

Zone number: N

Control horizon: $\mathcal{T} = \{1, 2, \dots\}$

Zone temperature: $\mathbf{T}_t = [T_t^1, \dots T_t^N]^{\top}$

Control variable: $\mathbf{a}_t = \left[\dot{m}_t^1, \dots \dot{m}_t^N, T_t^{da}\right]^{\top} \in \mathcal{A} \subset \mathbb{R}^{N+1}$

Objective weights: $\mathbf{w}_t = \left[w_t^{\mathcal{D}}, w_t^{\mathcal{E}}, w_t^{\mathcal{V}}\right]$, $\quad \mathbf{w}_t^{\top}\mathbf{1} = 1$

$\mathcal{D}(T_t^i)$: Building thermal discomfort of zone i at step t.

$\mathcal{E}_t$: HVAC energy consumption at step t. There is $\mathcal{E}_t \coloneqq P(\mathbf{a}_t, T_t^{out})\Delta t$

$\mathcal{V}_t$: Power limit violation penalty at step t.

$$\mathcal{V}_t \coloneqq \begin{cases} (P(\mathbf{a}_t, T_t^{out}) - \bar{P}_t)^2 & (P(\mathbf{a}_t, T_t^{out}) > \bar{P}_t) \\ 0 & (P(\mathbf{a}_t, T_t^{out}) \leq \bar{P}_t) \end{cases}$$

$\bar{P}_t$ is the DR limit issued by the utility, considering an incentive-based DR program.

Markov Decision Process Formulation

❖ State $\boldsymbol{s}_t = [\mathbf{T}_t, \mathbf{T}_{t,-K}^{out}, \mathbf{E}_t, \overline{\mathbf{P}}_t, t, \mathbf{w}_t] \in \mathcal{S}$.

✓ $\mathbf{T}_t \in \mathbb{R}^N$ represents zone temperatures.

✓ $\mathbf{T}_{t,-K}^{out} \in \mathbb{R}^K$ represents outdoor temperature for the last K steps.

✓ $\mathbf{E}_t = [f, sin_t, cos_t]$ are workday flag and sine/cosine representation of the time of the day.

✓ $\overline{\mathbf{P}}_t = [\overline{P}_t, \overline{P}_{t+1}, \dots, \overline{P}_{t+K-1}] \in \mathbb{R}^K$ indicates power limit for the next K steps.

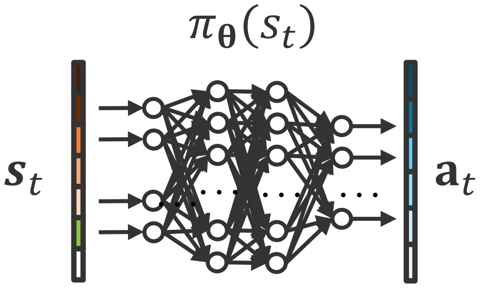

Fig. 2. RL policy network.

❖Action
$\mathbf{a}_t = [\dot{m}_t^1, \dot{m}_t^2, \dots, \dot{m}_t^N, T_t^{da}] \in \mathcal{A}$.

❖ Reward $r_t = -\mathbf{w}_t[\kappa_1 \sum_{i\in\mathcal{N}} \mathcal{D}(T_t^i), \kappa_2 \mathcal{E}_t, \kappa_3 \mathcal{V}_t]^\top$, negative value of single step cost.

RL Objective: $\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}}(\sum_{t\in\mathcal{T}} \gamma^t r_t)$

Approach: $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha \widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

$\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is the policy gradient estimated from sampled experience, e.g., using policy gradient theorem [4, Ch.13].

However, such policy search in RL typically amounts to solving non-convex optimization problems, converging to a poor-performing local optimum is likely, leading to unsatisfactory control performance.

Proposed Global-Local Policy Search

In order to achieve a faster convergence to a better policy, we propose combining complementary advantages from two different types of RL algorithms, letting them search the policy in two stages.

Stage I: Global Search

Using a *zero-order estimation (ZOE) based method* for policy gradient estimation [5]:

$$\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \frac{1}{\sigma} \mathbb{E}_{\epsilon \sim N(\mathbf{0}, \mathbf{I})} [\epsilon \cdot J(\boldsymbol{\theta} + \sigma\epsilon)]$$

- \+ Back-propagation (BP) free, fast gradient estimation.
- \+ Highly scalable.
- \+ Optimizing on the Gaussian smoothed objective, likely to avoid some poor-performing local optima.
- \- Inaccurate local convergence due to function smoothing.

Stage II: Local Tuning

Using a *policy gradient-based method* for policy gradient estimation (e.g., [6]):

$$\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathrm{t}}[\min(r_t(\boldsymbol{\theta})\hat{A}_t, clip(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon)\hat{A}_t]$$

- \+ Consider KL divergence during policy update (stable policy improvement).
- \+ Gradient-based learning on original objective gradient estimation (better local search ability).
- \- BP-based and conservative update (slower learning).
- \- Less scalable ($\mathcal{O}(N^2)$ communication complexity).
- \- Prone to be trapped in local optimum.

Combining these two types of RL algorithms allow us to leverage their strength, providing a faster convergence to a better local optimum.

Case Study [Experiment setup]

- Considering a five-zone benchmark building and minimize the daily cost $\left(\sum_{t\in\mathcal{T}} \mathbf{w}_t \left[\kappa_1 \sum_{i\in\mathcal{N}} \mathcal{D}(T_t^i), \kappa_2\mathcal{E}_t, \kappa_3\mathcal{V}_t\right]^{\top}\right)$ with control interval of 5-minute (i.e., $\mathcal{T} = \{1,2,\dots,288\}$).

- Building model for RL training is learned using data collected from EnergyPlus simulation.

- Exogenous data (e.g., outdoor temp) from July are used for training, and the trained RL controller will be test using unseen data from August.

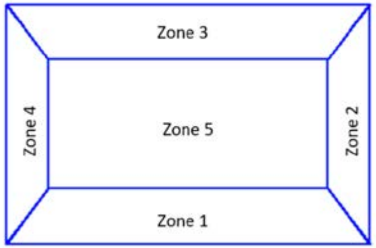

Fig. 3. Five-zone building investigated.

- RL controller training is implemented on the NREL high-performance computing (HPC) system.

  - ❑ Each computing node on NREL HPC system has dual 18-core processors with 96 GB memory [7].
  - ❑ For the Stage I training, we scale ES-RL [5], a ZOE-based RL algorithm, on 20 computing nodes, with a total of 684 rollout workers to sample control experience.
  - ❑ For the Stage II training, proximal policy optimization (PPO) algorithm [6] is used for policy fine-tuning. Single computing node is used as scaling PPO on multiple nodes does not bring significant benefit.

Case Study [Two-Stage RL Training]

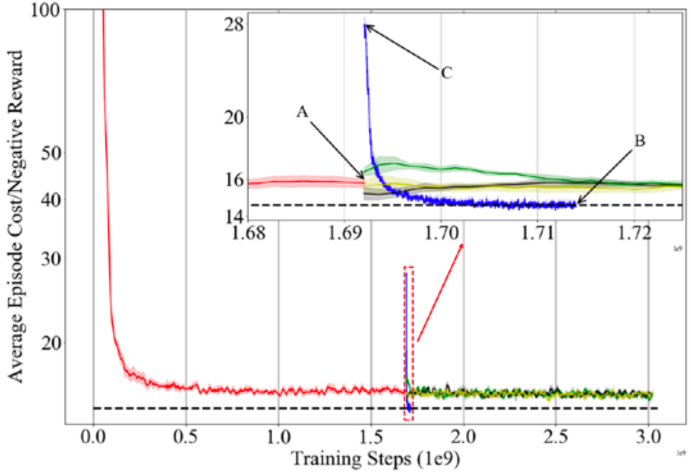

Fig. 4 Learning curves of the two-stage policy optimization. [Red]: Stage I ZOE-based global policy search; [Blue]: Stage II PG-based tuning; [Others]: unsuccessful ZOE-based tuning.

- Effectiveness of the two-stage learning.
- Using the ZOE-based method in Stage II for policy fine-tuning is not effective.

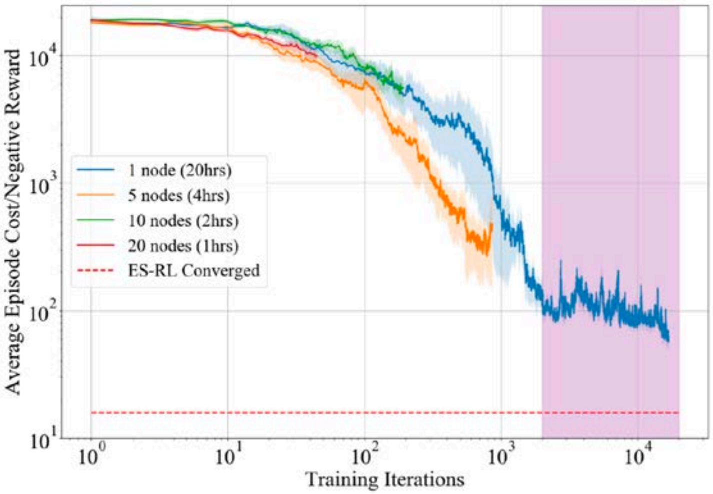

Fig. 5 Learning curves of using PG-based method (Proximal policy optimization (PPO) in this experiment).

- PPO training from scratch (without ES-RL) using the same computational resources.
- PPO training was not improved by naively scaling to multiple HPC computing nodes. See [8] for discussion
- Converged to local optima.

Case Study [Testing Scenarios]

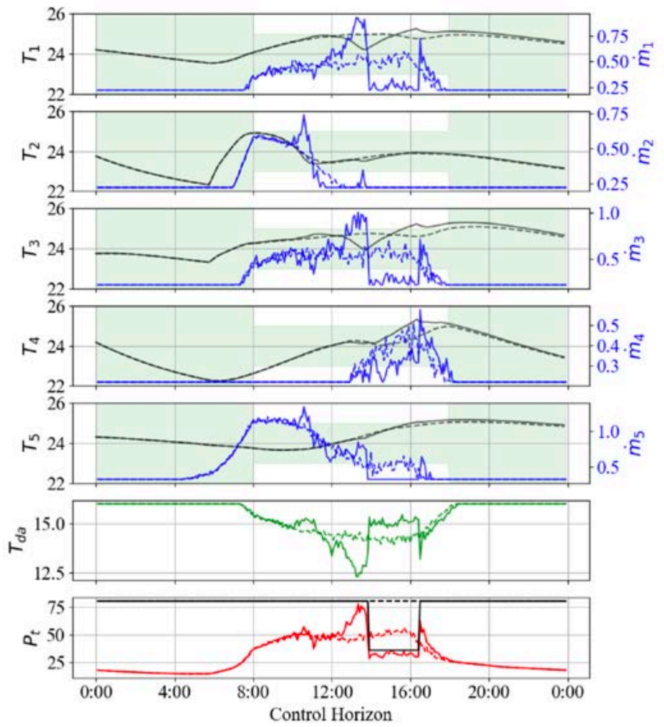

Fig. 6 Control demo of the trained RL controller in one testing day under DR scenario [solid] and non-DR scenario [dashed].

For one testing day, the control performance of the two-stage trained RL policy is shown under two scenarios:

- No DR event that day. (dashed lines)
- DR event (14:00-16:30) with 36 kW DR limit. (solid lines)

Performance of the global-locally searched policy:

- All zone temperature are mostly kept within the comfort band, except for some short period during DR events.
- Grid requirement can be successfully met.
- Proactive actions are taken to prepare the building for the incoming DR event.
- Though not explicitly instructed, the RL controller learned to differentiate different zone for better control.

TABLE I.
Comparison of Average Daily Cost of Test Scenarios

	Two-Stage RL	Linear MPC	Oracle MPC
DR	16.31	18.50	13.75
Non-DR	16.50	17.70	15.26

References

This presentation is based on the following publications:

[0-A] X. Zhang, R. Chintala, A. Bernstein, P. Graf, and X. Jin, “Grid-interactive multi-zone building control using reinforcement learning with global-local policy search,” in 2021 American Control Conference(ACC), May 25-28, 2021, pp. 4155–4162. Public link.
[0-B] X. Zhang, Y.Chen, A. Bernstein, R. Chintala, P. Graf, X. Jin, and D. Biagioni, “Two-stage reinforcement learning policy search for grid-interactive building control,” Under Review.

References appeared in previous slides:

[1] San Diego Gas & Electric, “AC Saver (Summer Saver),” Accessed: Nov. 16th, 2021. [Online]. Available: https://www.sdge.com/residential/savings-center/rebates/your-heating-cooling-systems/summer-saver-program.
[2] D. Kim, J. Braun, J. Cai, and D. Fugate, “Development and experimental demonstration of a plug-and-play multiple RTU coordination control algorithm for small/medium commercial buildings,” Energy and Buildings, vol. 107, pp. 279–293, Nov. 2015.
[3] X. Zhang, D. Biagioni, M. Cai, P. Graf, and S. Rahman, “An edge-cloud integrated solution for buildings demand response using reinforcement learning,” IEEE Transactions on Smart Grid, vol. 12, no. 1, pp. 420–431, Jan. 2021.
[4] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.
[5] Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, “Evolution strategies as a scalable alternative to reinforcement learning,”arXiv preprint arXiv:1703.03864, 2017.
[6] J . Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,”arXiv preprint arXiv:1707.06347,2017
[7] National Renewable Energy Laboratory, “Eagle System Configuration,” Accessed: Nov.16th, 2021. [Online]. Available: https://www.nrel.gov/hpc/eagle-system-configuration.html.
[8] S. McCandlish, J. Kaplan, D. Amodei, and OpenAI Dota Team, “An empirical model of large-batch training,” arXiv preprint arXiv:1812.06162, 2018.

Thank You!

Xiangyu.Zhang@nrel.gov

www.nrel.gov

NREL/PR-2C00-81534

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office.

NREL Transforming ENERGY