A Machine-Learning-Driven Energy-Efficient Resource Manager for Cloud Computing

Jinghua Wang[1], Asser Tantawi[2], Alaa S. Youssef[2], Tamar Eilam[2], Pradip Bose[2], Jiaxin Wan[1], Klara Nahrstedt[1], Deming Chen[1]

[1] UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

[2] IBM Research

BACKGROUND

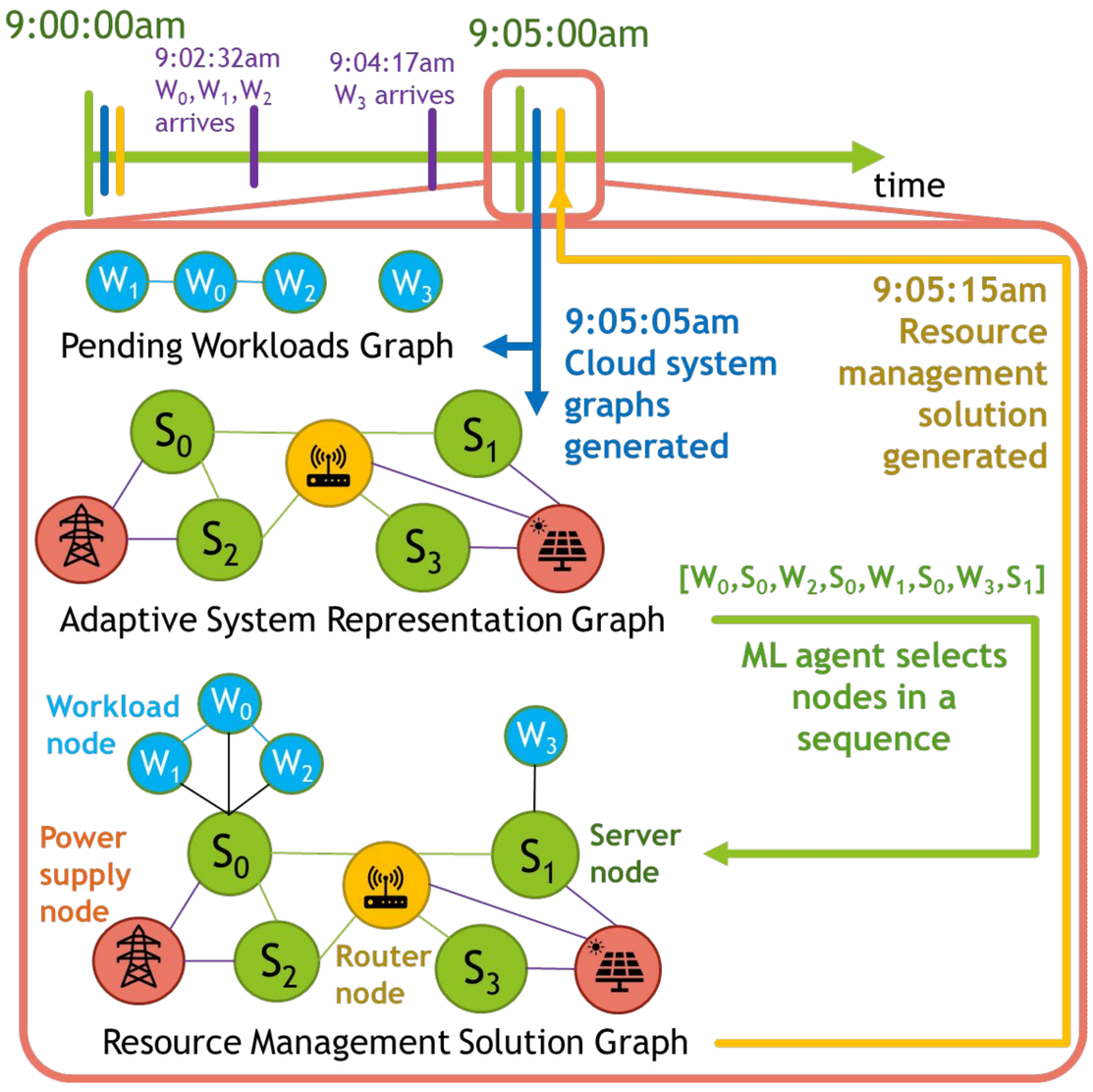
Importance of Energy-Efficient Cloud Computing

- Global data center electricity usage in 2021 was 220-320 TWh, around 0.9~1.3% of global final electricity demand [1]
- Training a Transformer neural network using neural architecture search results in CO_2 emission ~5 times than the total CO_2 emission in a car's lifetime [2]

INTRODUCTION

Energy-Efficient Resource Management in Cloud Computing

- **Scheduling**: which arrived workload should start running at each timestamp
- **Placement**: which server should the workload be placed onto
- How a good resource manager can save energy:
  - Place workloads on a server to utilize computing resources close to the most-energy-efficient utilization levels
  - Put server with no workloads to sleep
  - Place workloads to let their network traffics go though fewer physical links
  - Schedule more workloads when clean energy is highly available



A typical workflow of our ML-driven approach

CHALLENGES

Solving Global Optimal Resource Management Solution Is Computationally Expensive

- Even a simplified cloud resource management problem is NP-Hard

Modern Cloud Systems Are Dynamic And Scalable

- Each workload's computational needs fluctuate with time
- Each server's energy-efficient utilization levels vary with temperature
- Number of servers in the cloud system can change
- Clean energy availability changes with weather, time, etc.

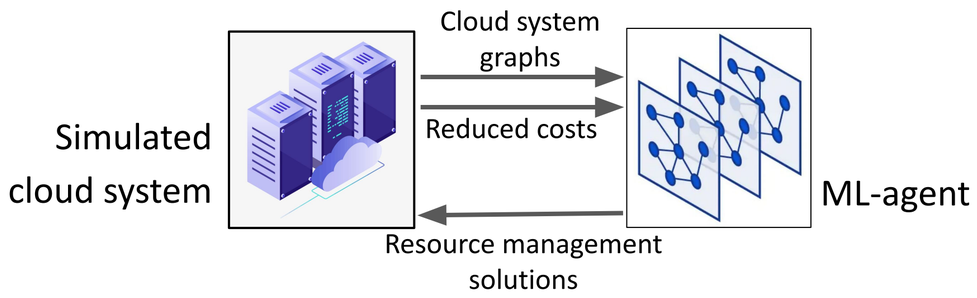
FUTURE WORK

- Generalize from simplified problem to non-simplified problem
- Generalize from cloud simulation to real-world cloud systems

METHODOLOGY

Simulation-Based Resource Manager Training

- Formulate a **cost function** from energy consumptions and service level objective (SLO) violations
- Use **graph neural networks** in ML agent to handle scalable cloud system representation graphs
- **Scale** resource management solutions by picking workload and server vertices in a sequence



Reinforcement learning setup of our ML-driven approach

PRELIMINARY RESULTS

Simplified cloud resource management problem:

- Simplified cloud resource management problem: each interval is a stand-alone problem
- Baseline resource manager: greedy algorithm
- Experiment setup: At each pod-server size, there are 1000 synthetic intervals for training, another 400 synthetic intervals for testing

$$\text{Test approximation ratio} = \frac{\sum_{i=1}^{400} \text{cost(Agent's solution to problem i)}}{\sum_{i=1}^{400} \text{cost(CPLEX's solution to problem i)}}$$

#servers \ #workloads	5	10	15	20
5	4.923	2.424	1.601	1.436
10	9.522	6.321	5.955	3.258
15	14.36	7.506	8.410	7.023
20	16.38	8.547	9.393	7.796

Test approximation ratios of greedy agent

#servers \ #workloads	5	10	15	20
5	1.202	1.155	1.047	1.029
10	1.213	1.242	1.296	1.190
15	1.234	1.208	1.330	1.345
20	1.266	1.246	1.363	1.287

Test approximation ratios of ML-agent

- ML-agent generates resource management solutions with cost reduction of 5.72x
- 5.72x energy reduction from greedy resource manager if successfully generated to real-world cloud systems

CONCLUSION

Preliminary results show that our ML-driven resource manager generates significantly better results than greedy algorithms on simplified problems in various scales. This demonstrates its potential of generating more energy-efficient solutions than existing methods and being more scalable in real-world cloud systems.

ACKNOWLEDGEMENT

This work is supported by **IBM-Illinois Discovery Accelerator Institute**.

REFERENCES

[1] G. Kamiya, "Data Centres and data transmission networks – analysis," *IEA (2022)*, Sep. 2022. [Online]. Available: https://www.iea.org/reports/data-centres-and-data-transmission-networks.

[2] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for Deep Learning in NLP," arXiv.org, Jun. 05, 2019. [Online]. Available: https://arxiv.org/abs/1906.02243.

Contact: jinghua3@illinois.edu

NAE Regional Meeting 2023