Towards Optimal Resource Utilization in Data Centers using Long Short-Term Memory

Kundjanasith Thonglek[1], Kohei Ichikawa[1],
Keichi Takahashi[1], Chawanat Nakasan[2], Hajimu Iida[1]

[1]Nara Institute of Science and Technology, Nara, Japan
[2]Kanazawa University, Kanazawa, Japan

Outline

- Introduction
- Methodology
- Experimental Result
- Conclusion

Introduction

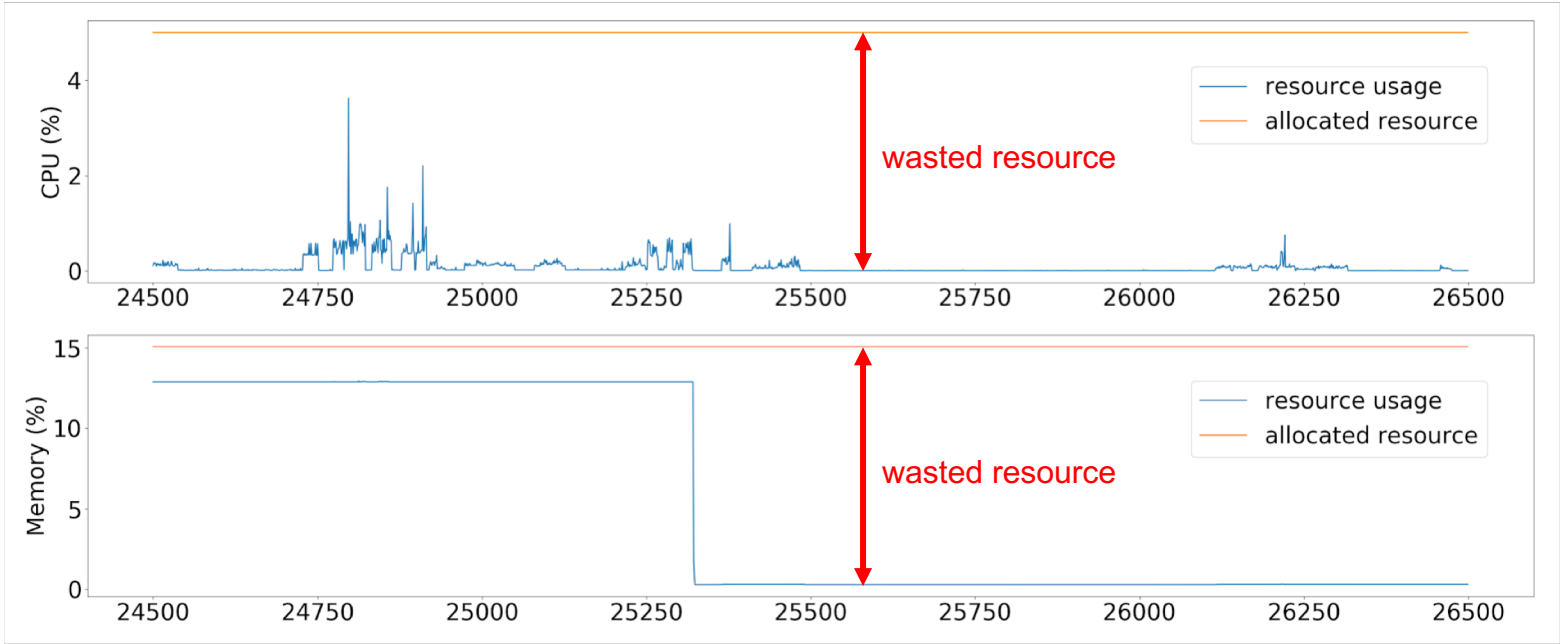
Introduction

Data Centers are centralized resources where computing and networking equipment is concentrated for the several purpose applications that handle large amounts of data and computing efficiently.

Data center consolidation requires a large number of well-managed migrations within a short period of time.

- Reduce Costs
- High Availability
- More Security
- Continuous Migration
- Compliance Audit
- Energy Efficiency
- Resource Utilization

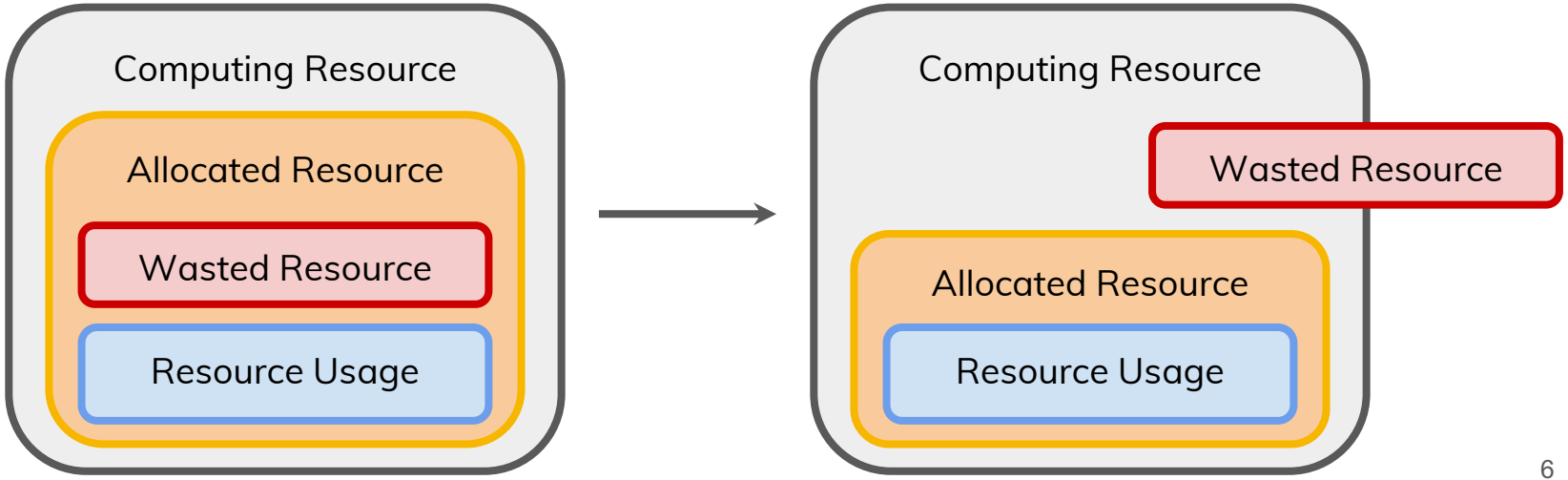
Resource Utilization



Resource Utilization is measuring how effectively the available computing resources to allocate their resources and usage with the less wasted resources.

Problem Statement

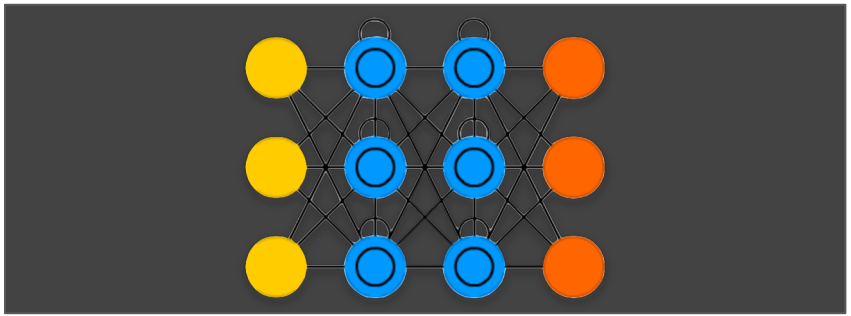
What is the suitable amount of allocated computing resources which can be allocated to complete the tasks with less wasted resources by the real workload in Google's data center ?



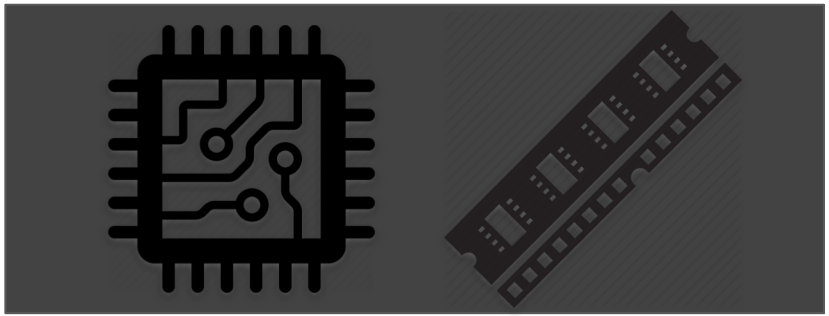
Objectives

We will optimize resource utilization in data center using Long Short-Term Memory to predict suitable allocated resource for increasing utilization rate.

The work applied Long Short-Term Memory technique to predict time series computing resources.



The work optimized computing resources utilization include CPU and memory in Data Center.

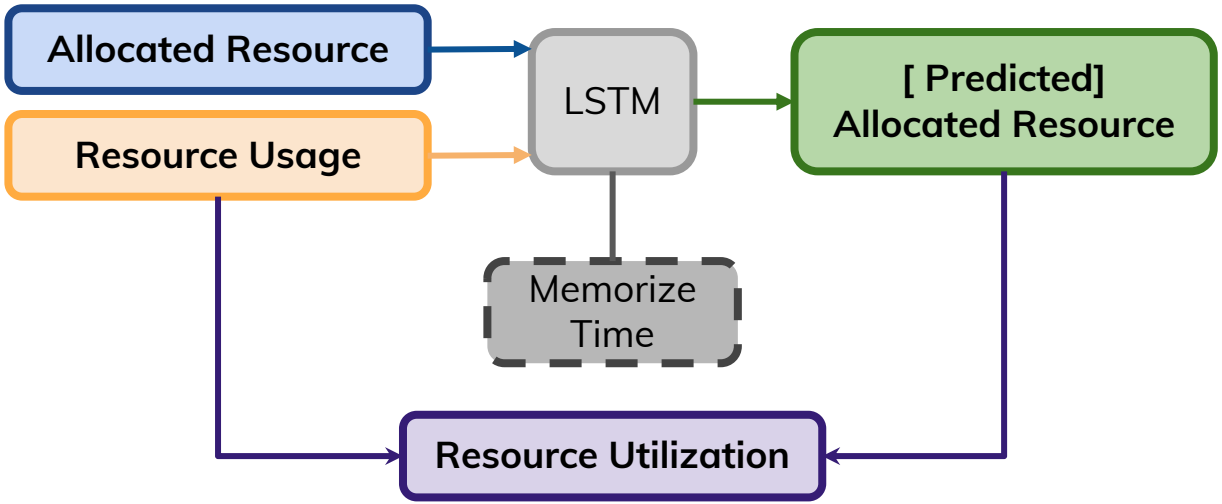


Approaches

Long Short-Term Memory or LSTM is applied to predict the allocated resources for minimize wasted resources with increasing the resource utilization rate.

Memorize time is the significant parameters to recognize the time interval in the memory cell of LSTM

Varying memorize time is depend on maximum and minimum execution time of the jobs in data center.



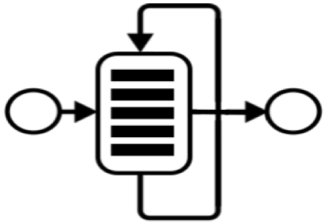
Methodology

Methodology

Towards Optimal Resource Utilization in Data Centers using LSTM

Google's Cluster Data

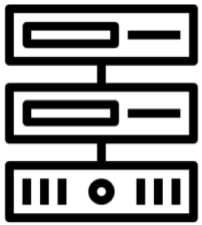
- Download and analyse Google's cluster usage data which is real workload in data center



Optimize Resource

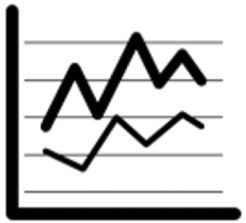
- Apply our LSTM model to inference with real workload in Google's data center by varying memorize time

Google's Open Data → **Long Short-Term Memory** → **Optimize Resource** → **Usage Simulation**



Long Short-Term Memory

- Design and implement LSTM to optimize allocated resource for increase resource utilization rate

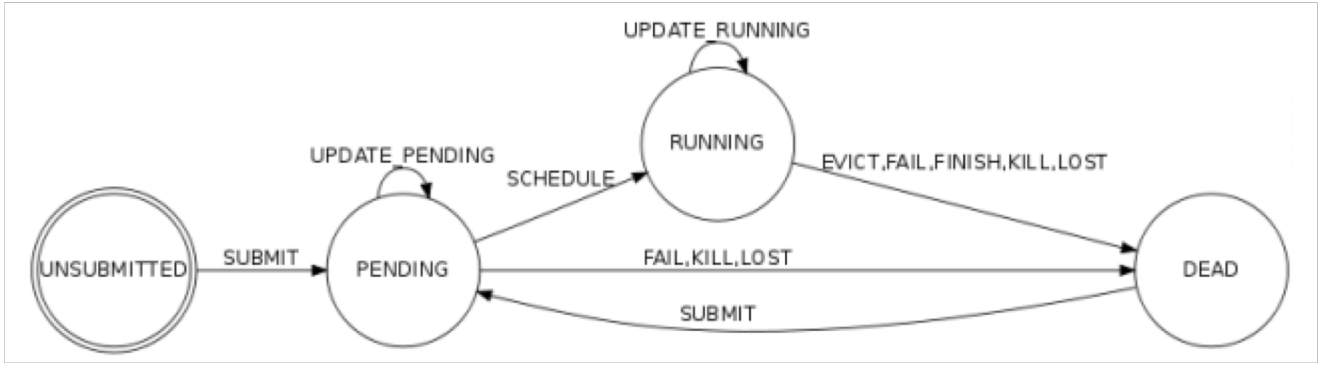


Usage Simulation

- Simulate resource utilization in data center using allocated resources and resources usage

Google’s cluster usage data

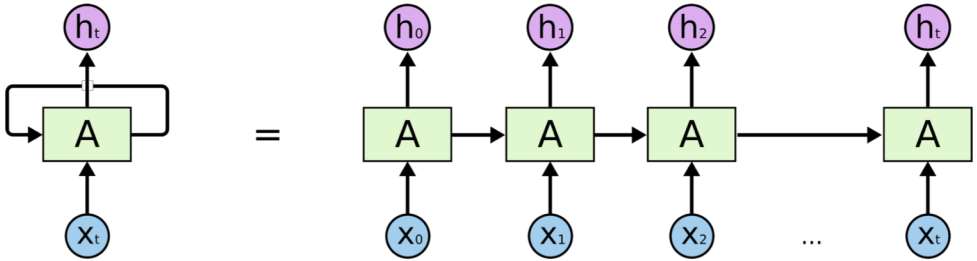
Google’s cluster usage data is real workload data in Google’s data center.



Computing Resource	Allocated Resource	Resource Usage
CPU	Allocated CPU	CPU usage
Memory	Allocated memory	memory usage

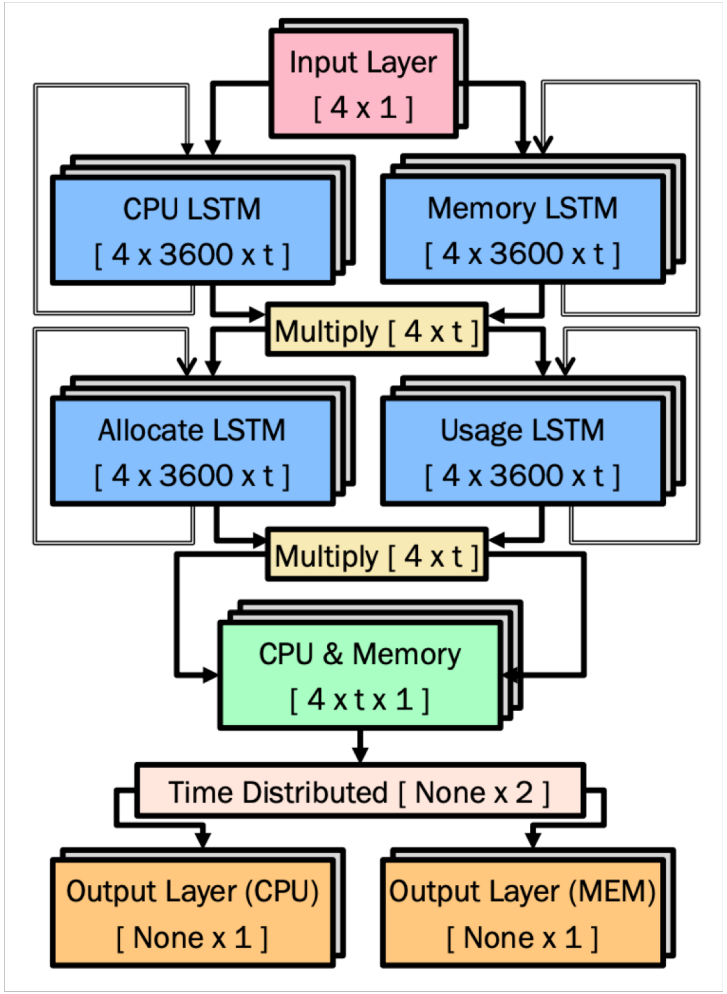
Long Short-Term Memory

Long Short-Term Memory or LSTM introduces long-term memory into recurrent neural networks.

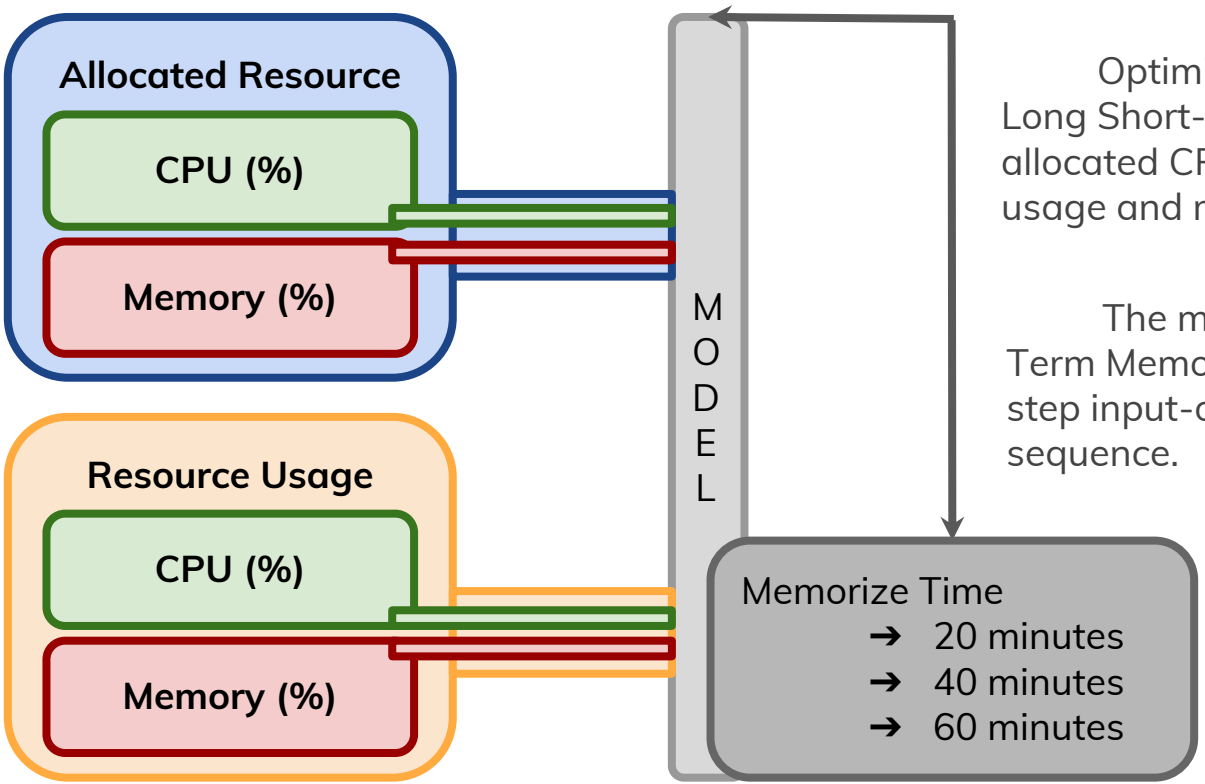


It migrates the vanishing gradient problem, which is where the neural network stops learning because the updates to the various weights within a given neural network become smaller and smaller.

There are two states that are being transferred to the next cell; the cell state and the hidden state.



Optimize Resource

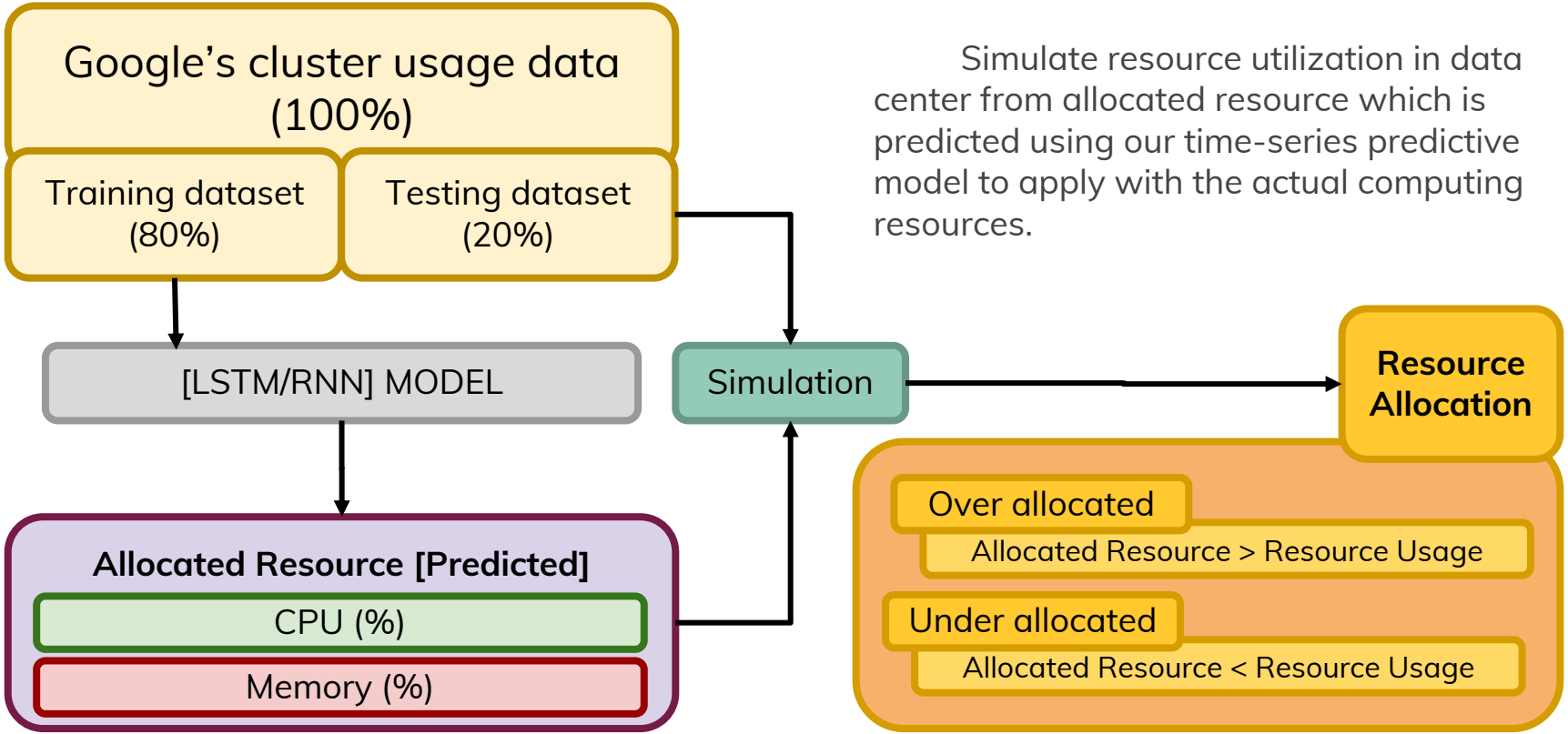


Optimizing resources by implement Long Short-Term Memory model using allocated CPU, allocated memory, CPU usage and memory usage.

The memorize time in Long Short - Term Memory model is memorizing each step input-output pair of values in each sequence.

Varying memorize time in Long Short-Term Memory model to predict allocated resources.

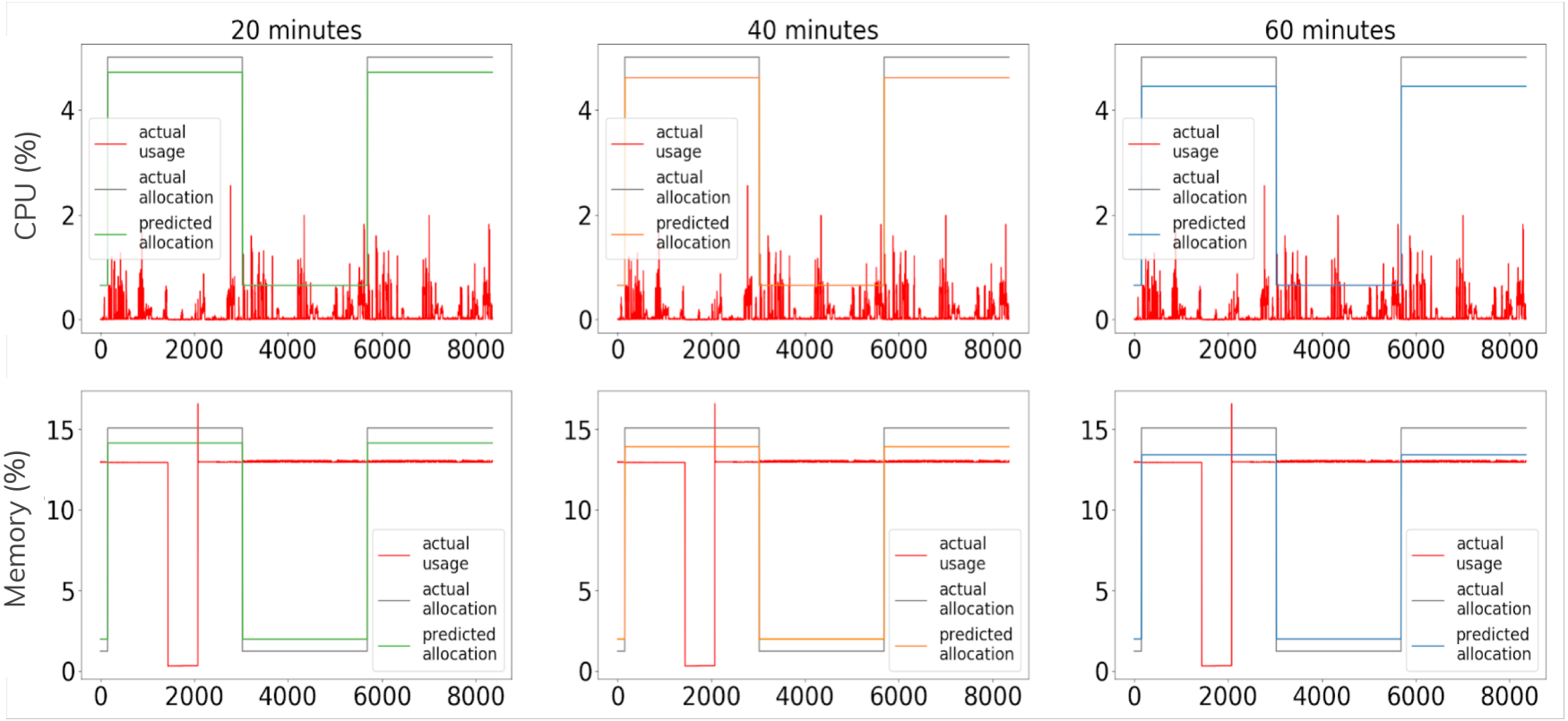
Usage Simulation



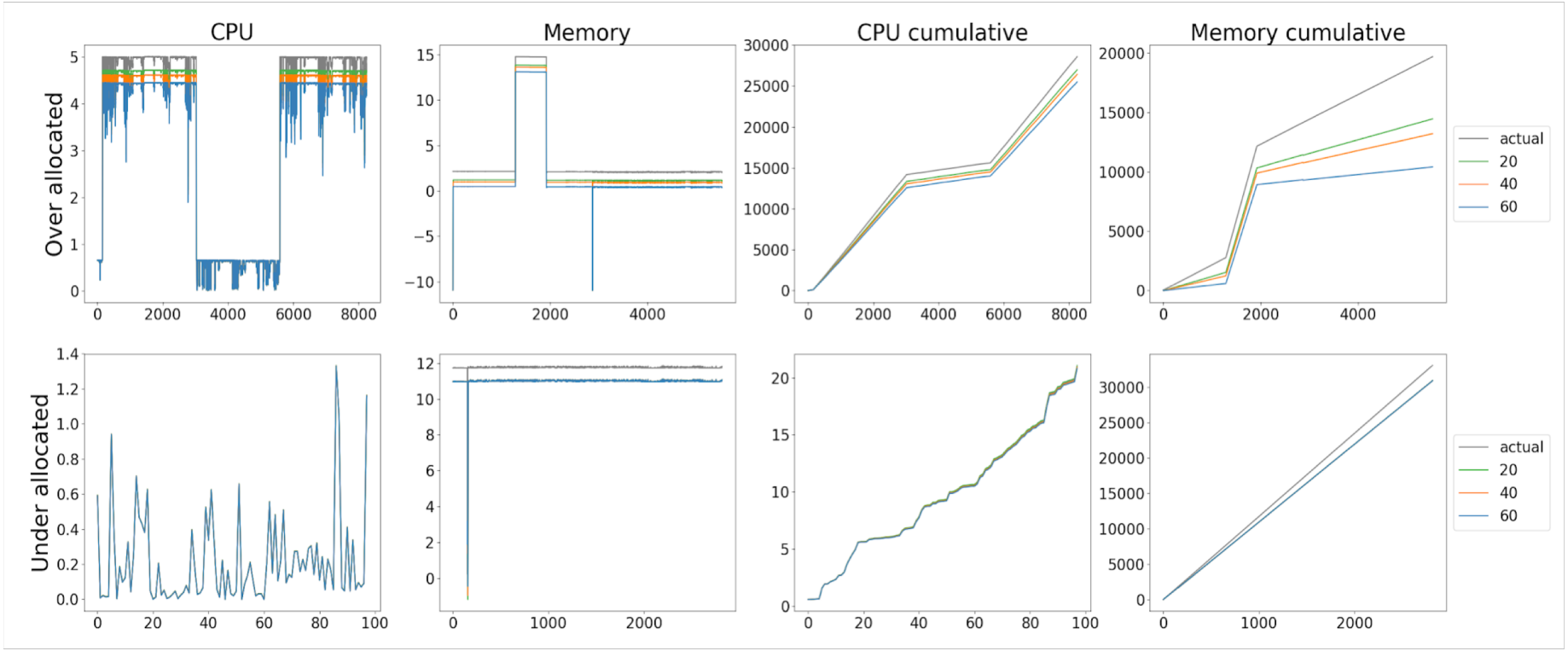
Simulate resource utilization in data center from allocated resource which is predicted using our time-series predictive model to apply with the actual computing resources.

Result

Allocated Resources

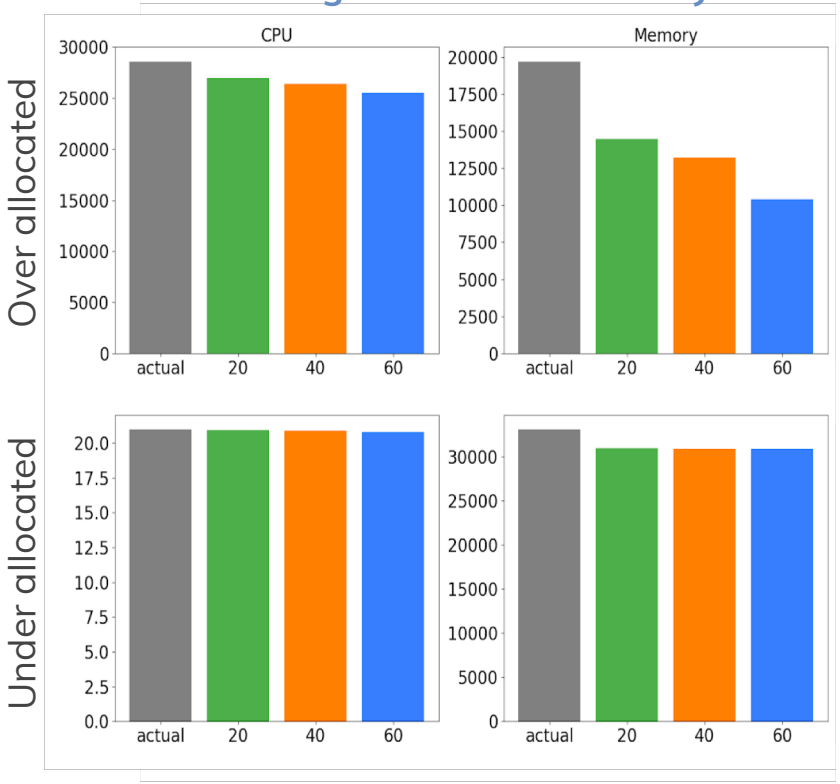


Wasted Resources

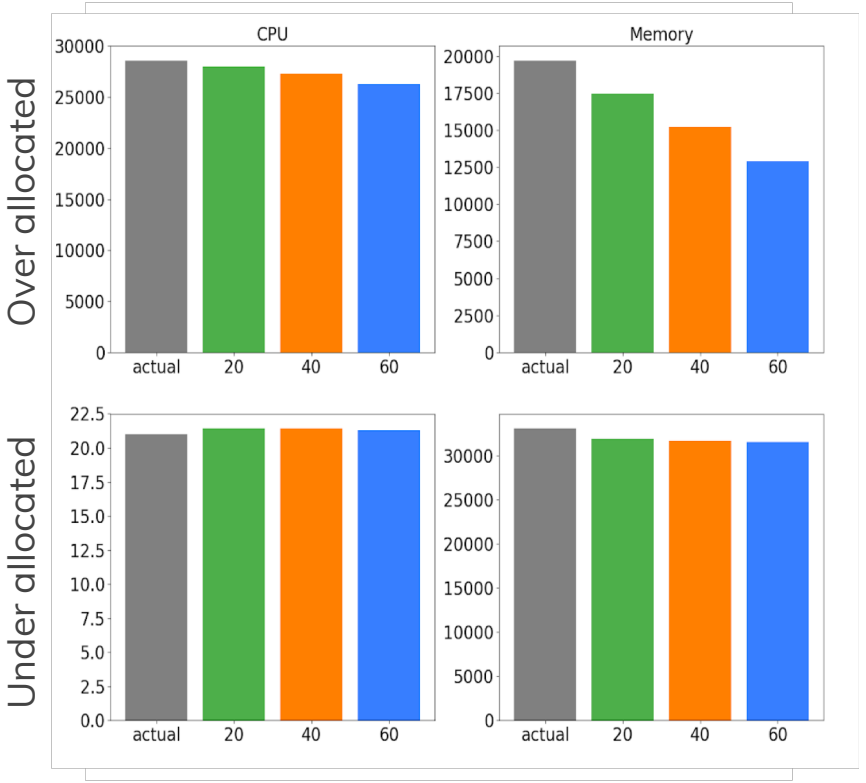


LSTM & RNN

Long Short-Term Memory



Recurrent Neural Network



Conclusion

Conclusion

- We study how to optimize the resource utilization in data center using Long Short-Term Memory
  - Discovered the impact from various memorize time in Long Short-Term Memory model.
  - Analyzed the real workload include allocated resource and resource usage in Google's data center.
  - Improved the suitable allocated resources to increase resource utilization rate.
- We would like to apply the other time-series forecasting techniques to optimize the resource utilization.

Email: thonglek.kundjanasith.ti7@is.naist.jp

Software Design & Analysis Laboratory, NAIST