ENERGY-EFFICIENT CLOUD TASK SCHEDULING USING DEEP LEARNING

K. D. S. A. Chandrasiri

248054V

Master of Science (Major Component Research)

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

April 2026

ENERGY-EFFICIENT CLOUD TASK SCHEDULING USING DEEP LEARNING

K. D. S. A. Chandrasiri

248054V

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science (Major Component Research)

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

April 2026

DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 2026-04-22

The supervisor should certify the Thesis with the following declaration.

The above candidate has carried out research for the Master of Science (Major Component Research) Thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Prof. Dulani Meedeniya

Signature of the Supervisor: Date: 22/04/2026

DEDICATION

I dedicate this thesis to my parents, whose unconditional love, guidance, and sacrifices have been the foundation of my educational journey and personal growth. I am deeply grateful to my beloved wife for her constant encouragement, patience, and understanding, which sustained me throughout the challenges of this work.

I also extend my sincere appreciation to my supervisor, Prof. Dulani Meedeniya, for her invaluable guidance, insightful feedback, and continuous support throughout the completion of this thesis. Her mentorship has been instrumental in shaping my research and academic development.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Prof. Dulani Meedeniya, for her invaluable guidance, continuous encouragement, and insightful feedback throughout the course of this research. Her expertise and support have been instrumental in shaping this work and in strengthening my academic and research skills.

I am also grateful to the academic and administrative staff of the department for their support and assistance during my postgraduate studies. Their contributions have created a conducive environment for learning and research.

Finally, I would like to thank my family and friends for their unwavering support and encouragement, which have been a source of strength and motivation throughout this journey.

ABSTRACT

This thesis presents a deep learning-based approach for energy-efficient cloud workflow scheduling by modeling scheduling as a multi-objective reinforcement learning problem that simultaneously optimizes makespan, energy consumption, and Quality of Service (QoS) . The work first introduces GNN-Flat, a baseline Graph Neural Network (GNN) scheduler that represents workflow dependencies using Directed Acyclic Graph (DAG) structures and demonstrates the feasibility of applying graph-based learning to cloud task scheduling. Building on insights from this implementation, the research proposes 2SD-GAT, a two-stage deep reinforcement learning scheduler based on Graph Attention Networks (GAT) , which forms the core contribution of the thesis. The 2SD-GAT architecture separates task selection and resource allocation decisions while leveraging attention mechanisms to capture inter-task dependencies and preference-based reward optimization, enabling improved Pareto trade-offs across competing objectives. Extensive experiments using synthetic and real-world datasets show superior performance compared with heuristic and learning-based baselines. The results show that the proposed approach achieves a 26.8% hypervolume gain with a 4.5x lower Inverted Generational Distance (IGD), along with a makespan improvement of 14.08%. Finally, the research bridges theory and practice by developing a pluggable RL-based scheduling framework, integrations with Apache Airflow and Kubernetes, and supporting tools including workflow engines, execution components, and a web-based GUI, demonstrating how the proposed scheduler can be deployed in realistic cloud orchestration environments.

Keywords: cloud, cloud-task-scheduling, deep-reinforcement-learning, energy-consumption, multi-objective optimization

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	**Description**
ACO	Ant Colony Optimization
AI	Artificial Intelligence
CAGR	Compound Annual Growth Rate
CIS	Cloud Information Service
CLI	Command Line Interface
CP	Constraint programming
CRDs	Custom Resource Definitions
CSO	Cat Swarm Optimization
DAG	Directed Acyclic Graph
DDQN	Double Deep Q-Network
DFJSSP	Dynamic FJSSP
DRL	Deep Reinforcement Learning
DS	Deep Set
DVFS	Dynamic Voltage and Frequency Scaling
EA	Evolutionary Algorithms
EDF	Earliest Deadline First
ETL	Extract-Transform-Load
FCFS	First-Come-First-Serve
FJSSP	Flexible Job Shop Scheduling Problem
GA	Genetic Algorithms
GAT	Graph Attention Networks
GIN	Graph Isomorphism Network
GNN	Graph Neural Network
GSA	Gravitational Search Algorithm
HEFT	Heterogeneous Earliest Finish Time
HV	Hypervolume
IGD	Inverted Generational Distance
ILP	Integer Linear Programming
KinD	Kubernetes in Docker
LSTM	Long Short-Term Memory
MDP	Markov Decision Process
ML	Machine Learning
MLP	Multi-Layer Perceptron

Abbreviation	**Description**
MOHEFT	Multi-Objective Heterogeneous Earliest Finish Time
MOO	Multi-Objective Optimization
MTBF	Mean Time Between Failures
NIST	National Institute of Standards and Technology
PPO	Proximal Policy Optimization
PSO	Particle Swarm Optimization
QoS	Quality of Service
RL	Reinforcement Learning
RNNs	Recurrent Neural Networks
RR	Round Robin
SPEC	Standard Performance Evaluation Corporation
VM	Virtual Machine

LIST OF APPENDICES