Abstract

Cloud computing has emerged as the backbone of scalable and cost-effective computational infrastructures, particularly for scientific and data-intensive applications. Within this paradigm, task scheduling plays a critical role in mapping tasks onto virtualized resources while optimizing multiple objectives such as execution time, monetary cost, energy consumption, and resource utilization. The challenge, however, lies in the intrinsic complexity of the task scheduling problem, which is NP-hard due to the combinatorial nature of resource-task allocations, availability of cloud resources, and the necessity to meet Quality of Service constraints like deadlines and budgets. This thesis addresses this multifaceted problem by developing five novel AI-enhanced approaches, each targeting a different dimension of the task scheduling problem. These methods are evaluated using well-known scientific workflows, such as Montage, CyberShake, SIPHT, Inspiral, and Epigenomics, and randomly generated tasks with varying characteristics under different resource configurations. The evaluation is multi-objective in nature and grounded in rigorous performance metrics and statistical validation.

The first chapter of the thesis introduces the core motivations, challenges, and research objectives. It elaborates on the complications of dealing with large-scale workloads in heterogeneous and elastic cloud environments. It models the system architecture for the task and workflow scheduling problem and provides the mathematical formulations for task model, workflow model, resource model, task scheduling model that incorporates makespan and scheduling time and workflow scheduling model that incorporates makespan, cost, energy consumption and resource utilization models. Further, it discusses the frameworks such as CloudSim, WorkflowSim and Google Colaboratory, as the simulation tools utilized in this work. The chapter discusses the experimental methodology employed to validate the proposed approaches. It discusses the benchmark workflows on which the scheduling approaches were tested, followed by the metrics used for

performance analysis of the approaches, including the Pareto-optimality metrics and statistical analysis. It outlines the objectives of the research, which include developing scalable, deadline and budget-constrained scheduling, local optima avoidance-based scheduling and reinforcement learning-based methods. The introduction also presents a high-level overview of the proposed approaches.

Chapter 2 presents an extensive literature survey that categorizes existing workflow scheduling algorithms into several major paradigms: rule-based heuristics, metaheuristic algorithms, hybrid approaches and machine learning-based techniques. This chapter critiques the limitations of mono-objective and static scheduling algorithms, including their inability to handle task dependencies, fluctuating resource availability, and competing optimization objectives. It also explores hybrid models and emerging reinforcement learning techniques and lists their shortcomings in terms of convergence speed, generalizability, and constraint handling. The survey concludes by identifying significant research gaps: the need for hybrid, adaptive, multi-objective and scalable algorithms that can provide feasible, near-optimal schedules for tasks with varying characteristics while simultaneously satisfying deadline and budget constraints.

Chapter 3 addresses this gap with the first two proposed algorithms. The Deadline and Budget-Constrained Archimedes Optimization Algorithm extends the physics-inspired Archimedes Optimization Algorithm to the workflow scheduling domain by embedding budget and deadline constraints into the optimization process. ADB dynamically adjusts the search process based on the feasibility and fitness of candidate schedules, offering improved convergence behavior, incorporating exploration and exploitation and better constraint compliance than other approaches. The chapter then introduces the Multi-Layer Perceptron Optimized Archimedes approach, which hybridizes the AOA metaheuristic with a Multi-Layer Perceptron. The MLP serves as a surrogate evaluator that learns from historical schedule data to predict optimal control parameter values. By using the fitness based on previous scheduling decisions, it enables more effective tuning of the AOA's parameters, thereby guiding the search process more efficiently. This hybrid approach accelerates convergence while maintaining high schedule quality and constraint satisfaction. Experimental results show that both ADB and MLPOA outperform baseline methods on various performance indicators.

Chapter 4 continues the exploration of metaheuristic enhancements with two additional con-

tributions. The Modified Local Escaping Archimedes Optimization Algorithm is proposed to overcome local optima entrapment and enhance exploration in high-dimensional scheduling problems. MLEAO incorporates a Local Escaping Operator that is selectively activated to introduce perturbation into the population. The second model in this chapter, Manta Ray Foraging Optimization with Opposition-Based Learning, draws inspiration from manta ray foraging behavior and augments it with various Opposition-Based Learning strategies. These opposition techniques are designed to widen the search space coverage and accelerate convergence towards global optima while maintaining population diversity. The model retains the LEO mechanism from MLEAO to ensure perturbation in the stages of stagnation. The combined strategies make MRFOBL highly effective in solving complex, multi-objective scheduling problems. Comparative experiments demonstrate that these approaches produce superior Pareto fronts and outperform other multi-objective optimization algorithms while maintaining computational efficiency.

Chapter 5 introduces a paradigm shift from metaheuristic-based optimization to adaptive decision-making through reinforcement learning. The Advantage Actor-Critic Strategy models the scheduling problem as a Markov Decision Process and learns a policy to assign tasks to virtual machines in a way that maximizes cumulative reward. The actor network proposes actions, while the critic network computes the advantage and evaluates those actions. The reward function is carefully designed to balance multiple objectives: minimizing makespan and scheduling time. Unlike static optimization approaches, A2CS continuously learns and adapts to environmental changes, including fluctuating resource availability and dynamic task arrivals. This makes it particularly suitable for real-time scheduling scenarios. The experimental evaluation shows that A2CS generalizes well across different task characteristics and maintains high-quality performance. It represents an important contribution to the development of autonomous, self-learning scheduling systems in cloud environments.

Each of the five proposed approaches is rigorously evaluated using scientific workflows and tasks of varying sizes and characteristics. The thesis employs a rich set of performance metrics, makespan, execution cost, energy consumption, resource utilization, as well as Pareto-optimality indicators like hypervolume, S-metric and dominance rate. Statistical significance is validated using t-tests and ANOVA to ensure the robustness of performance comparisons. Collectively, the models demonstrate clear advantages over existing scheduling algorithms in both constrained and unconstrained scenarios. The ADB and MLPOA models provide robust performance in

constrained environments, while MLEAO and MRFOBL enhance solution diversity and convergence quality. A2CS, on the other hand, introduces adaptability and self-learning capabilities, positioning itself as a scalable solution for real-time, cloud-native environments.

The work addresses key challenges such as multi-objective trade-offs, deadline and budget constraint handling, scalability, search diversity, local optima avoidance and adaptability. The research findings open new avenues for future exploration, including online adaptive scheduling, federated reinforcement learning, explainable AI in scheduling, and integration with emerging paradigms like edge computing and fog computing. This thesis thus lays a solid foundation for the development of next-generation intelligent task schedulers capable of meeting the diverse and evolving demands of cloud computing environments.

Keywords: Task Scheduling, Workflow Scheduling, Cloud Computing, Multi-Objective Optimization, Deadline Constraint, Budget Constraint, Metaheuristic and Hybrid Algorithms, Archimedes Optimization Algorithm, Manta Ray Foraging Optimization, Opposition Based Learning, Local Escaping Operator, Multi-Layer Perceptron, Reinforcement Learning, Advantage Actor-Critic Learning, Scientific Workflows, Makespan Optimization, Cost Minimization, Energy Efficiency, Resource Utilization, AI-enhanced Scheduling, Pareto Optimality.