МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

**"САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО"**

Институт компьютерных наук и технологий

Направление **02.03.01** : Математика и компьютерные науки

Literature Review:

«Modern Scheduling Algorithms in Cloud-Fog and High-Performance Computing Environments»

Student: Darya Yashnova 5130201/20002

Teacher: Ph.D, Motorin D. E.

Saint-Petersburg, 2024

1 Introduction

Cloud computing is a reliable and scalable way to use computers. Cloud services receive requests to perform work, which can be simple tasks or more complex workflows.

Task planning consists of distributing work between computers to make sure that it is performed quickly and efficiently. This includes taking into account factors such as cost, power consumption, and how reliable the security system is.

Cloud computing consumes a lot of energy, which is harmful to the environment. Thus, cloud service providers need to find ways to use less energy.

Previous studies have looked at different ways to plan tasks and ways to improve the efficiency of this system. Different methods and approaches were used in these studies.

Recently, some new methods have been developed using artificial intelligence (AI). These methods allow you to automatically find the best way to schedule tasks without requiring much human help. They showed promising results in solving the problem of task scheduling in cloud computing.

There are many different strategies that can be used to improve energy efficiency in data centers. These strategies can be grouped into different categories depending on where they work (location, infrastructure, hardware, or software). Task scheduling is a type of software optimization that helps you manage resources. There are also many heuristic methods that solve these problems. In this review, you can look at the results of some research in this area.

2 Analysis

2.1 Heuristic approaches

Heuristic algorithms are often used to solve the task scheduling optimization problem.

The WOA(Whale Optimization Algorithm), a metaheuristic algorithm inspired by the hunting behavior of humpback whales, is often used to optimize task scheduling algorithms [1], [2], [3].

The PSO(Particle Swarm Optimization) heuristic is also used — an algorithm based on swarm behavior that searches for the optimal solution by moving in a multi-dimensional solution space with a swarm of particles [1], [4], [2], [5]. Some algorithms, such as WOA and GMPSO, face problems predicting future tasks, which may limit their use in situations with dynamic changes in load [3],[5]. Fitness functions are often used to evaluate the effectiveness of optimization algorithms[1], [2], [3], [5].

2.2 DRL and expert policies

DRL (Deep Reinforcement Learning) is an algorithm that automatically selects the optimal policy. Its main idea is to obtain rewards through interaction between the agent and the environment in order to maximize returns and achieve specific optimization goals [6], [2]. Several DRL algorithms based on expert knowledge have been developed [6], [2]. DRL can optimize task scheduling in conjunction with WOA [2] and through a network of experts [6]. In particular, the improved DQN algorithm has been applied to develop an energy-efficient task scheduler that also takes into account response time and average running time [2], [7], [8], [9]. Algorithms based on deep neural networks and DQN show higher efficiency in terms of reducing response time and energy consumption compared to traditional methods. The approaches that use actor-critic algorithms focus on tasks in a multi-cloud environment, where optimization focuses on reducing costs and execution time [6], [8], [7], [9].

2.3 Optimization Parameters

In addition to performance metrics such as runtime, cloud computing systems also face the challenge of minimizing power consumption. Several studies emphasize the importance of energy-efficient task optimization [1], [2], [3]. Some researches discuss methods and algorithms designed to reduce energy consumption while maintaining a high level of performance. Various algorithms evaluate metrics such as execution time [5], [10], [6], [3] or average waiting time [10], [6], cost [1], [4], [5]. Optimization of tasks in cloud computing often involves balancing competing goals, so often several performance parameters are evaluated at once, all parameters can be seen on Table 1.

2.4 Prioritization of tasks

Some studies use dynamic prioritization of tasks, for example, the Enhanced Shortest Job First algorithm[4], which focuses on task completion time and resource usage. While others [6], apply deep learning and generative adversarial networks (GANS) to optimize resource allocation. Some studies have a prioritization based on task's size [10], [3], [5].

2.5 Test environments

Different studies have been conducted in different environments such as CloudSim [10],[5], Python simulations [4], [8], and high-performance computing (HPC) [6],[7]. This variety of test environments can affect the overall performance of algorithms in real-world conditions. As can be seen on Table 2 these algorithms are often tested on the Cybershake, Montage, Epigenomics, Sipht, and Inspiral streams [1], [5], [9].

2.6 Results

Studies demonstrate the effectiveness of various novel algorithms in resource allocation and task scheduling, consistently outperforming existing approaches across multiple metrics. Modified SJF scheduling [10] shows a significant reduction in average waiting time (20-40%), highlighting its efficiency in resource allocation. Improvements in total execution cost (TEC) are also reported, with reductions ranging from 8.64% to 22.68% compared to PSO and 30.43% to 71.31% compared to standard WOA [1]. Resilience testing [4] reveals the superiority of Swarm2 (100% resilience), with other algorithms exhibiting varying degrees of resilience (90-60%). GARLSched demonstrates superior performance in average waiting time (AVGwt) and average blocking time (AVGbsld) metrics across all workloads, achieving 41-66% improvement under high load [6]. DWOA exhibits significant advantages in convergence speed, solution quality, and stability, confirmed by Wilcoxon and Friedman tests [2]. While a proposed algorithm minimizes execution time, migration, and power consumption [3], it lacks predictability of upcoming tasks. Focusing on high-performance workflows, GMPSO shows negligible gains for small workflows but adapts well to various sizes [5]. DeepMIC consistently outperforms greedy algorithms, reducing average delay by up to 25.03% and average task response time by up to 20.75% [7]. EETS balances energy consumption and response time, demonstrating optimal solution finding and outperforming other methods across multiple metrics [8]. Finally, MCWS-A3C achieves significant execution time reduction while maintaining high resource utilization (above 60%) and adapts well to varying workloads [9]. Overall, these studies highlight the potential for algorithmic advancements to optimize resource utilization, task scheduling, and resilience in diverse computing environments.

3 Conclusion

Optimizing task scheduling in cloud computing is a multi-faceted task that requires a comprehensive approach. Heuristic algorithms, such as WOA and PSO, demonstrate high efficiency in solving time planning problems. DRL algorithms based on expert knowledge offer innovative approaches to automatically selecting the optimal security policy. Energy efficiency is also an important aspect that requires special attention. Research in this area is ongoing, and further developments can lead to significant improvements in the performance and efficiency of cloud systems.

References

[1] S. Bansal, H. Aggarwal, A hybrid particle whale optimization algorithm with application to workflow scheduling in cloud–fog environment, Decision Analytics Journal 9 (2023) 100361. doi:https://doi.org/10.1016/j.dajour.2023.100361.

[2] T. Shu, Z. Pan, Z. Ding, Z. Zu, Resource scheduling optimization for industrial operating system using deep reinforcement learning and woa algorithm, Expert Systems with Applications 255 (2024) 124765. doi:https://doi.org/10.1016/j.eswa.2024.124765.

[3] S. Mangalampalli, G. R. Karri, G. N. Satish, Efficient workflow scheduling algorithm in cloud computing using whale optimization, Procedia Computer Science 218 (2023) 1936–1945, international Conference on Machine Learning and Data Engineering. doi:https://doi.org/10.1016/j.procs.2023.01.170.

[4] B. Dupont, N. Mejri, G. Da Costa, Energy-aware scheduling of malleable hpc applications using a particle swarm optimised greedy algorithm, Sustainable Computing: Informatics and Systems 28 (2020) 100447. doi:https://doi.org/10.1016/j.suscom.2020.100447.

[5] H. Hafsi, H. Gharsellaoui, S. Bouamama, Genetically-modified multi-objective particle swarm optimization approach for high-performance computing workflow scheduling, Applied Soft Computing 122 (2022) 108791. doi:https://doi.org/10.1016/j.asoc.2022.108791.

[6] J. Li, X. Zhang, J. Wei, Z. Ji, Z. Wei, Garlsched: Generative adversarial deep reinforcement learning task scheduling optimization for large-scale high performance computing systems, Future Generation Computer Systems 135 (2022) 259–269. doi:https://doi.org/10.1016/j.future.2022.04.032.

[7] X. Pei, P. Sun, Y. Hu, D. Li, L. Tian, Z. Li, Multi-resource interleaving for task scheduling in cloud-edge system by deep reinforcement learning, Future Generation Computer Systems 160 (2024) 522–536. doi:https://doi.org/10.1016/j.future.2024.06.033.

[8] H. Hou, A. Ismail, Eets: An energy-efficient task scheduler in cloud computing based on improved dqn algorithm, Journal of King Saud University - Computer and Information Sciences 36 (8) (2024) 102177. doi:https://doi.org/10.1016/j.jksuci.2024.102177.

[9] X. Tang, F. Liu, B. Wang, D. Xu, J. Jiang, Q. Wu, C. P. Chen, Workflow scheduling based on asynchronous advantage actor–critic algorithm in multi-cloud environment, Expert Systems with Applications 258 (2024) 125245. doi:https://doi.org/10.1016/j.eswa.2024.125245.

[10] Y. Pachipala, K. S. Sureddy, A. S. Kaitepalli, N. Pagadala, S. S. Nalabothu, M. Iniganti, Optimizing task scheduling in cloud computing: An enhanced shortest job first algorithm, Procedia Computer Science 233 (2024) 604–613, 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024). doi:https://doi.org/10.1016/j.procs.2024.03.250.

Appendix

Table 1 : Comparative Study for Task Scheduling methods Optimization in High-Performance Computing

	Heuristic methods				Math methods								
Research Paper	GA	WOA	PSO	Prioritization	SJF	Fitness Function	MDP	GAN	Greedy	Actor-critic	DRL	Comparative studies	Optimization parameters
Optimizing Task Scheduling in Cloud Computing: An Enhanced Shortest Job First Algorithm				Dynamic prioritization	+							+	Task completion time, Resource utilization
A Hybrid Particle Whale Optimization Algorithm with application to workflow scheduling in cloud–fog environment		+	+			+						+	TET, TEC
Energy-aware scheduling of malleable HPC applications using a Particle Swarm optimised greedy algorithm			+	FIFO					+			+	Task Reorganization parameters, Server Shutdown Options
GARLSched: Generative adversarial deep reinforcement learning task scheduling optimization for largescale high performance computing systems							+	+		+	+	+	AVGwt, AVGbsld
Resource scheduling optimization for industrial operating system using deep reinforcement learning and WOA algorithm		+		Service metrics							+	+	Convergence speed, solution quality, performance stability
Efficient Workflow Scheduling algorithm in cloud computing using Whale Optimization		+		Task's size, execution time		+						+	Migrationtime, energy consumption, makespan
Genetically-modified Multiobjective Particle Swarm Optimization approach for high-performance computing workflow scheduling	+		+	Time and cost to complete the task		+						+	Inverted Generational Distance (IGD), Hypervolume (HV)
Multi-resource interleaving for task scheduling in cloud-edge system by deep reinforcement learning				Normalized weight factor, minimizing the weighted-sum delay penalty			+			+	+	+	Flow Completion Time, Average Computing Waiting Time, Task Response Time
EETS: An energy-efficient task scheduler in cloud computing based on improved DQN algorithm				Prioritized Experience Replay			+			+	+	+	Average Task Response Time, Makespan, Average Work Time, Energy Consumption
Workflow scheduling based on asynchronous advantage actor–critic algorithm in multi-cloud environment				based on a reward function			+			+	+	+	Makespan, Cost

Table 2 : Testing results

		Testing environment				Compared algorithms												
Research Paper	Testing flows	CloudSim	Workflowsim	CEC2017 test suite	Other environments	SJF	GWO	FCFS	MLP	WOA	PWOA	PSO	FIFO	Rand-Param	Garlsched	DRL	DQN	GA
Optimizing Task Scheduling in Cloud Computing: An Enhanced Shortest Job First Algorithm	«Light», «Moderate» and «Heavy» workloads	+				SJF, MSJF												
A Hybrid Particle Whale Optimization Algorithm with application to workflow scheduling in cloud–fog environment	Cybershake, Montage, Epigenomics, Sipht and Inspiral		+							+	+	+						
Energy-aware scheduling of malleable HPC applications using a Particle Swarm optimised greedy algorithm					Simulation environment was developed using python							Swarm1, Swarm2, Swarm3	FIFO, FIFO-rRcfg, FIFO-Poff	RandParam1, RandParam2, RandParam3				
GARLSched: Generative adversarial deep reinforcement learning task scheduling optimization for largescale high performance computing systems	Lublin-256, HPC2N, SDSC-BLUE, SDSC-SP2				Intel (R) Xeon (R) Gold 5218 CPU @ 2.30 GHz				+						+	+		
Resource scheduling optimization for industrial operating system using deep reinforcement learning and WOA algorithm				+			+			EWOA, WOA, BOA		+				RL-GWO, RLWOA		+
Efficient Workflow Scheduling algorithm in cloud computing using Whale Optimization			+							Pwhale		+						+
Genetically-modified Multiobjective Particle Swarm Optimization approach for high-performance computing workflow scheduling	Cybershake, Montage, Epigenomics, Sipht and Inspiral	+										SMPSO, OMOPSO, GMPSO						
Multi-resource interleaving for task scheduling in cloud-edge system by deep reinforcement learning					2683 v3 CPU, a GTX 2080Ti graphics card, and an Ubuntu 18.04.1 system											SD-NNC, CCEC, DECO, Deep MIC		
EETS: An energy-efficient task scheduler in cloud computing based on improved DQN algorithm	Alibaba Cluster Traces v2018				Intel(R) i7-12700H (2.3 GHz) processor with 32G RAM and NVIDIA 3060 (6G) GPU												EETS, DQN, DDQN	
Workflow scheduling based on asynchronous advantage actor–critic algorithm in multi-cloud environment	Montage, CyberShake, Epigenomics, LIGO, Sipht				Intel i7-7700K CPU and an NVIDIA GTX 1070 graphics card (8 GB graphics RAM) GPU												HEFT, ACO, SDQN, MCWS-A3C	