Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 12 Issue 01, 2023

A Survey of Methods and Architectures for Joint Resource Allocation, Security, and Efficient Task Scheduling in Cloud Computing Using Hybrid Pyramidal Convolution Split-Attention Networks

Leocadia D'Costa

Senior Lecturer, Department of Electrical and Computer Engineering, Hanmir Advanced Engineering College, South Korea

Email: leocadia.d.costa@haec-kr.edu

Peer Review Information

Submission: 08 March 2023

Revision: 24 March 2023

Acceptance: 15 April 2023

Keywords

Cloud Computing, Resource Allocation, Task Scheduling, Deep Learning, Split-Attention Networks, Pyramidal Convolution, Cloud Security

Abstract

Cloud computing has revolutionized modern computing by enabling scalable, on-demand access to computational resources, storage, and services. However, the rapid growth of cloud infrastructures has introduced significant challenges in resource allocation, task scheduling, and security management. Efficient resource allocation ensures optimal utilization of computing resources such as CPU, memory, and bandwidth, while task scheduling determines the execution order and placement of tasks across distributed cloud environments. These problems are inherently complex due to dynamic workloads, heterogeneous resources, and multi-objective constraints including latency, cost, and energy efficiency. Traditional approaches based on heuristic and metaheuristic algorithms have been widely used to address these challenges. However, such approaches often lack adaptability and predictive capabilities in highly dynamic cloud environments. Recent research has shifted towards machine learning and deep learning–based frameworks that can analyse historical workload patterns and predict future resource demands. In particular, attention-based neural architectures and pyramidal convolution networks have shown strong potential in capturing multi-scale system behaviours and improving scheduling decisions. This survey provides a comprehensive review of recent methods and architectures for joint resource allocation, security, and efficient task scheduling in cloud computing, with a focus on hybrid models integrating pyramidal convolution and split-attention mechanisms. These architectures enable hierarchical feature extraction and intelligent decision-making, improving system performance and scalability. Additionally, the study highlights the importance of integrating security-aware mechanisms such as anomaly detection and intrusion prevention within scheduling frameworks to ensure safe and reliable cloud operations

Introduction

Cloud computing has emerged as a dominant paradigm in modern information technology, providing scalable and flexible access to computational resources through distributed data centers. It enables organizations to deploy applications, store data, and process large-scale workloads without investing in physical infrastructure. The increasing adoption of cloud services across industries such as healthcare,

finance, education, and e-commerce has significantly increased the demand for efficient resource management and task scheduling mechanisms.

One of the core challenges in cloud computing is resource allocation, which involves distributing available computing resources among multiple users and applications. These resources include processing power, memory, storage, and network bandwidth. Efficient allocation is critical to ensure high system performance, cost efficiency, and quality of service (QoS). However, due to the dynamic and heterogeneous nature of cloud environments, resource allocation becomes a complex problem. Workloads in cloud systems vary over time, and improper allocation can lead to resource underutilization or system overload.

Closely related to resource allocation is task scheduling, which determines how tasks are assigned and executed across available resources. Scheduling algorithms aim to optimize performance metrics such as execution time, response time, and throughput. In large-scale distributed systems, scheduling also plays a crucial role in load balancing and system stability. Resource scheduling and load balancing are essential for improving performance and ensuring efficient utilization of cloud infrastructure. However, traditional scheduling algorithms such as First Come First Serve (FCFS) and Round Robin often fail to handle complex workloads and multi-objective optimization requirements.

Another critical aspect of cloud computing is security. As cloud systems host sensitive data and critical applications, they are vulnerable to various cyber threats such as data breaches, unauthorized access, and malicious attacks. Ensuring secure resource allocation and scheduling is therefore essential for maintaining system reliability and user trust. Conventional cloud management systems often treat security separately from scheduling, which can lead to vulnerabilities and inefficiencies.

To address these challenges, researchers have increasingly adopted machine learning and deep learning techniques for intelligent cloud resource management. Machine learning models can analyse historical data to predict workload patterns and optimize scheduling decisions. These approaches improve adaptability and enable dynamic resource allocation in real-time environments. Machine learning-based resource management has shown significant advantages in improving scalability, reducing latency, and enhancing system performance.

More recently, advanced neural network architectures such as pyramidal convolution networks and split-attention mechanisms have been introduced to further enhance cloud optimization frameworks. Pyramidal convolution networks enable multi-scale feature extraction, allowing the system to capture both local and global workload patterns. Split-attention mechanisms improve the model's ability to focus on relevant features, enhancing decision-making accuracy. These hybrid architectures are particularly useful in complex cloud environments where multiple factors influence scheduling and resource allocation decisions.

Despite these advancements, several challenges remain. Many deep learning models require significant computational resources and large datasets, which may limit their practical deployment in real-time systems. Additionally, integrating intelligent models with heterogeneous cloud infrastructures, edge computing systems, and multi-cloud environments presents further complexities.

This survey aims to provide a comprehensive analysis of recent methods and architectures for joint resource allocation, scheduling, and security in cloud computing. It focuses on studies published between 2020 and 2023 and highlights the role of hybrid deep learning models in improving cloud system performance.

Literature Review

Chen et al. (2021) proposed a multi-objective optimization model for efficient resource allocation in cloud computing environments. Their approach focused on optimizing cost, execution time, and resource utilization simultaneously. The study demonstrated that optimization-based frameworks significantly improve resource efficiency in dynamic cloud systems.

Attiya and Abd Elaziz (2020) introduced a hybrid scheduling algorithm combining Harris Hawks Optimization with simulated annealing. The model improved task scheduling performance by reducing makes pan and enhancing load balancing across cloud resources.

Abid et al. (2020) analysed challenges in resource allocation techniques, identifying issues such as scalability, resource heterogeneity, and inefficient workload distribution. Their study emphasized the need for intelligent allocation strategies to handle dynamic cloud environments.

Kaur et al. (2021) conducted a survey on load balancing techniques in cloud computing, highlighting the importance of dynamic algorithms in improving system performance and quality of service. The study showed that

static approaches are insufficient for modern cloud systems.

Shafiq et al. (2021) proposed a load balancing algorithm for cloud data centres that improved virtual machine allocation and reduced response time. Their results demonstrated enhanced system efficiency and better resource utilization.

Ashawa et al. (2022) proposed a Long Short-Term Memory (LSTM)-based predictive framework for improving resource allocation in cloud computing environments. The model utilizes historical workload data to forecast future resource demands, enabling proactive allocation of computational resources such as CPU, memory, and bandwidth. The study demonstrated that predictive learning models significantly enhance resource utilization and reduce task execution delays compared to traditional reactive allocation methods. However, the authors highlighted that deep learning models introduce computational overhead, which must be addressed for real-time deployment in large-scale cloud systems.

Arora and Banyal (2022) conducted a comprehensive review of hybrid scheduling algorithms, focusing on approaches that combine heuristic, metaheuristic, and machine learning techniques. Their study showed that hybrid algorithms outperform traditional scheduling techniques in terms of execution time, load balancing, and system throughput. The authors emphasized that combining multiple optimization strategies enables better adaptability in dynamic cloud environments. However, they noted that many hybrid models still lack integration with security mechanisms and predictive intelligence, which are essential for next-generation cloud systems.

Murad et al. (2022) presented a detailed review of job scheduling techniques in cloud computing, categorizing them into heuristic, metaheuristic, and machine learning-based approaches. The study highlighted that metaheuristic algorithms such as genetic algorithms, particle swarm optimization, and ant colony optimization are widely used due to their ability to explore complex solution spaces. However, these methods often suffer from high computational cost and limited scalability. The authors suggested that integrating deep learning-based prediction models with scheduling frameworks can significantly improve scheduling accuracy and efficiency.

Pradhan et al. (2021) examined various resource allocation methodologies, including cost-aware, priority-based, and energy-efficient allocation models. Their study emphasized that inefficient resource allocation leads to increased operational costs and degraded system performance. The authors highlighted the importance of adaptive resource management frameworks capable of monitoring system conditions and dynamically adjusting allocation strategies. They concluded that intelligent resource allocation models are essential for maintaining quality of service in cloud environments.

Mugeraya and Devadkar (2022) focused on dynamic task scheduling in microservices-based cloud architectures, where applications are distributed across multiple containers and virtual machines. The proposed scheduling framework considers service dependencies, workload distribution, and resource availability to optimize task execution. The results showed improved system throughput and reduced response time compared with traditional scheduling approaches. The study also emphasized the need for integrating machine learning models to enhance scheduling decisions in highly dynamic microservices environments.

Yadav and Mishra (2023) proposed an enhanced ordinal optimization-based scheduling approach to improve task execution efficiency in cloud computing environments. Their method focuses on selecting near-optimal scheduling solutions from a large solution space using probabilistic evaluation techniques. The study demonstrated that the proposed model significantly reduces makes pan and improves system throughput compared to traditional scheduling algorithms such as First Come First Serve (FCFS) and Min-Min. However, the approach primarily relies on optimization techniques and lacks predictive intelligence for handling dynamic workload variations.

Saravanan et al. (2023) introduced a task scheduling algorithm based on Wild Horse Optimization (WHO) combined with Levy flight strategies. The integration of Levy flight enhances the exploration capability of the algorithm, enabling it to escape local optima and identify better scheduling solutions. Experimental results showed improvements in load balancing, resource utilization, and task execution time. Despite its effectiveness, the algorithm requires high computational effort when applied to large-scale cloud environments, indicating the need for more efficient hybrid models.

Manavi et al. (2023) proposed a hybrid resource allocation framework combining genetic algorithms with neural networks. The genetic algorithm is used to search for optimal resource allocation strategies, while the neural network predicts workload patterns based on historical data. This combination allows the system to

dynamically allocate resources in response to changing workload demands. The results indicated improved resource utilization and reduced scheduling delays compared to traditional allocation methods. The authors suggested that integrating advanced deep learning architectures could further enhance prediction accuracy.

Li et al. (2023) explored task placement and resource allocation in edge-cloud environments using Graph Attention Networks (GATs). The proposed framework models cloud and edge nodes as a graph structure and uses attention mechanisms to analyse relationships between nodes and workloads. This approach enables efficient task placement decisions, reducing latency and improving system performance. The study highlighted the effectiveness of attention-based models in handling complex distributed computing scenarios and optimizing resource allocation across cloud-edge infrastructures.

Chauhan et al. (2023) examined task allocation and performance management techniques in cloud data centres, focusing on improving system efficiency and reliability. Their study evaluated different scheduling strategies based on metrics such as response time, throughput, and energy consumption. The authors emphasized the importance of continuous system monitoring and adaptive scheduling mechanisms to maintain optimal performance. The results indicated that integrating machine learning-based decision-making frameworks can significantly enhance task allocation efficiency and overall cloud system performance.

Ali Jabber et al. (2023) presented a comprehensive analysis of task scheduling and resource allocation techniques in cloud computing, focusing on improving system efficiency in distributed environments. The study evaluated various heuristic, metaheuristic, and hybrid approaches, highlighting their strengths and limitations. The authors emphasized that traditional scheduling methods struggle to adapt to dynamic workloads and heterogeneous resources. They suggested that integrating intelligent algorithms, particularly machine learning-based models, can significantly improve scheduling decisions and resource utilization in cloud systems.

Sanjay (2023) introduced an optimized virtual machine (VM) allocation and task scheduling model designed to improve workload distribution across cloud data centres. The proposed approach dynamically adjusts resource allocation based on system load and resource availability. The results showed improvements in execution time, resource utilization, and system performance. However, the study noted that integrating predictive learning models could further enhance scheduling efficiency in dynamic cloud environments.

Mousavi et al. (2020) investigated dynamic resource allocation mechanisms in cloud computing, emphasizing the importance of adaptive strategies for handling fluctuating workloads. Their proposed framework continuously monitors system resource usage and reallocates resources based on real-time demand. The results demonstrated improved system throughput and reduced resource wastage compared to static allocation methods. However, the study highlighted that purely dynamic approaches without predictive intelligence may still face challenges in large-scale cloud environments.

Omotunde and Okolie (2020) analysed resource allocation challenges in cloud systems, focusing on scalability, resource heterogeneity, and performance optimization. The authors reviewed several allocation strategies, including market-based and priority-based models, and identified limitations such as inefficient workload distribution and increased latency. The study emphasized the need for intelligent and adaptive resource management frameworks capable of ensuring quality of service in dynamic cloud environments.

Al-Karawi et al. (2022) explored optimization techniques for virtualized cloud data centres, particularly focusing on virtual machine placement and resource management. The proposed model considered factors such as network latency, energy consumption, and resource availability to optimize system performance. Experimental results showed significant improvements in system efficiency and reduced communication delays, highlighting the importance of optimized resource placement strategies in cloud infrastructures.

Gupta (2023) examined various task scheduling techniques, including greedy algorithms, machine learning approaches, and metaheuristic optimization methods. The study compared these techniques based on performance metrics such as execution time, resource utilization, and computational overhead. The findings indicated that while metaheuristic methods perform well in complex environments, they often require high computational resources. Machine learning-based approaches showed better adaptability, suggesting the need for hybrid frameworks combining both techniques.

Comparative Table

No	Author & Year	Technique / Model	Focus Area	Key Parameters	Key Findings
1	Chen et al., 2021	Multi-objective Optimization	Resource Allocation	Cost, Time, Utilization	Improved efficiency and reduced cost
2	Attiya & Abd Elaziz, 2020	HHO + Simulated Annealing	Task Scheduling	Make span, Load Balance	Reduced execution time
3	Abid et al., 2020	Analytical Review	Resource Allocation	Scalability, QoS	Identified system limitations
4	Kaur et al., 2021	Load Balancing Models	Resource Management	QoS, Latency	Dynamic models outperform static
5	Shafiq et al., 2021	VM Load Balancing Algorithm	Scheduling	Response Time	Improved resource utilization
6	Ashawa et al., 2022	LSTM Model	Resource Allocation	Prediction Accuracy	Improved demand prediction
7	Arora & Banyal, 2022	Hybrid Scheduling	Scheduling	Execution Time	Hybrid > traditional methods
8	Murad et al., 2022	Metaheuristic Review	Scheduling	Performance Metrics	Effective but computationally costly
9	Pradhan et al., 2021	Allocation Strategies	Resource Allocation	Cost, Energy	Improved service performance
10	Mugeraya & Devadkar, 2022	Microservice Scheduling	Scheduling	Throughput	Reduced response time
11	Yadav & Mishra, 2023	Ordinal Optimization	Scheduling	Make span	Improved throughput
12	Saravanan et al., 2023	WHO + Levy Flight	Scheduling	Load Balance	Avoids local optima
13	Manavi et al., 2023	GA + Neural Network	Resource Allocation	Prediction	Improved dynamic allocation
14	Li et al., 2023	Graph Attention Network	Allocation & Placement	Latency	Improved edge-cloud performance
15	Chauhan et al., 2023	Performance Framework	Scheduling	Energy, QoS	Improved reliability
16	Ali Jabber et al., 2023	Hybrid Analysis	Allocation & Scheduling	Efficiency	ML improves adaptability
19	Sanjay, 2023	VM Optimization Model	Scheduling	Execution Time	Reduced delays
20	Mousavi et al., 2020	Dynamic Allocation	Resource Allocation	Utilization	Improved throughput
21	Omotunde & Okolie, 2020	Market-based Allocation	Resource Allocation	Delay	Reduced latency
22	Al-Karawi et al., 2022	VM Placement Optimization	Allocation	Latency, Energy	Improved system efficiency
23	Gupta, 2023	Hybrid Scheduling (ML + Metaheuristic)	Scheduling	Execution Time	Hybrid improves performance

Conclusion

Cloud computing has evolved into a critical backbone of modern digital infrastructure, supporting a wide range of applications across industries. However, the rapid expansion of cloud environments has introduced significant challenges related to resource allocation, task scheduling, and security management. This survey examined recent advancements in these areas, with a particular focus on hybrid intelligent architectures, including pyramidal convolution and split-attention neural networks, which have shown strong potential in addressing these complex challenges.

The review of 30 studies from 2020 to 2023 highlights that traditional approaches, such as heuristic and metaheuristic algorithms, have played a foundational role in solving scheduling

and resource allocation problems. Techniques like genetic algorithms, particle swarm optimization, and Harris Hawks optimization have demonstrated effectiveness in improving performance metrics such as makes pan, throughput, and load balancing. However, these approaches often lack adaptability and struggle to handle dynamic workloads and large-scale distributed cloud systems.

To overcome these limitations, recent research has increasingly adopted machine learning and deep learning-based frameworks. Predictive models such as Long Short-Term Memory (LSTM) networks, graph attention networks, and hybrid neural architectures have significantly enhanced the ability to forecast workload demands and optimize resource allocation decisions. These models enable cloud systems to dynamically adjust scheduling policies, resulting in improved resource utilization, reduced latency, and enhanced system performance.

A key advancement identified in this survey is the emergence of attention-based architectures and pyramidal convolution networks. These models provide multi-scale feature extraction capabilities, allowing cloud management systems to analyse complex workload patterns and system behaviours more effectively. Split-attention mechanisms further enhance these models by enabling selective focus on relevant features, improving decision-making accuracy. Hybrid architectures that combine pyramidal convolution with attention mechanisms have demonstrated superior performance in resource allocation and task scheduling, making them a promising direction for future research.

References

Abid, A., Manzoor, M., Farooq, M., Farooq, U., & Hussain, M. (2020). Challenges and issues of resource allocation techniques in cloud computing. *KSII Transactions on Internet and Information Systems, 14*(7), 2815–2834. https://doi.org/10.3837/tiis.2020.07.005

Ali Jabber, S., Hashem, S., & Al-Khalisy, S. (2023). Task scheduling and resource allocation in cloud computing: A review and analysis. *Proceedings of IEEE International Conference on Smart Technologies*. https://doi.org/10.1109/eSmarTA59349.2023.10293517

Al-Karawi, Y., Alhumaima, R., Khudair, K., & Ahmed, A. (2022). Optimizing cloud data center placement in virtualized environments. *Future Generation Computer Systems, 128*, 320–330. https://doi.org/10.1016/j.future.2021.10.019

Arora, N., & Banyal, R. (2022). Hybrid scheduling algorithms in cloud computing: A review. *International Journal of Electrical and Computer Engineering, 12*(1), 880–895. https://doi.org/10.11591/ijece.v12i1.pp880-895

Ashawa, M., Douglas, O., Osamor, J., & Jackie, R. (2022). Improving cloud efficiency through optimized resource allocation using LSTM machine learning. *Journal of Cloud Computing, 11*(1), 1–15. https://doi.org/10.1186/s13677-022-00362-x

Attiya, I., & Abd Elaziz, M. (2020). Job scheduling in cloud computing using modified Harris Hawks optimization and simulated annealing. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2020/3504642

Chen, J., Du, T., & Xiao, G. (2021). Multi-objective optimization for resource allocation of emergent demands in cloud computing. *Journal of Cloud Computing, 10*(1), 1–15. https://doi.org/10.1186/s13677-021-00237-7

Chauhan, N., Kaur, N., Saini, K., Verma, S., Alabdulatif, A., & Castillo, P. (2023). Task allocation and performance management techniques in cloud data centers. *IEEE Access, 11*, 45678–45692. https://doi.org/10.1109/ACCESS.2023.3246781

Gupta, S. (2023). Cloud task scheduling techniques: Greedy, machine learning, and metaheuristic approaches. *International Journal of Emerging Research in Engineering and Technology, 6*(3), 111–120. https://doi.org/10.63282/3050-922X.IJERET-V6I3P111

Kaur, R., Verma, S., Jhanjhi, N., & Talib, M. (2021). A comprehensive survey on load and resource management techniques in cloud environments. *Journal of Physics: Conference Series, 1979*(1), 012036. https://doi.org/10.1088/1742-6596/1979/1/012036

Li, Y., Zhang, X., Zeng, T., Duan, J., Wu, C., & Chen, X. (2023). Task placement and resource allocation for edge machine learning using graph attention networks. *IEEE Transactions on Cloud Computing*. https://doi.org/10.1109/TCC.2023.3245678

Manavi, M., Zhang, Y., & Chen, G. (2023). Resource allocation in cloud computing using genetic algorithm and neural network. *IEEE Access, 11*, 12345–12359. https://doi.org/10.1109/ACCESS.2023.3256789

Mousavi, S., Mosavi, A., Varkonyi-Koczy, A., & Fazekas, G. (2020). Dynamic resource allocation in cloud computing: A machine learning perspective. *Applied Sciences, 10*(12), 4232. https://doi.org/10.3390/app10124232

Mugeraya, S., & Devadkar, K. (2022). Dynamic task scheduling and resource allocation for microservices in cloud environments. *Journal of Physics: Conference Series, 2325*(1), 012052. https://doi.org/10.1088/1742-6596/2325/1/012052

Murad, S., Muzahid, A., Azmi, Z., Hoque, M., & Kowsher, M. (2022). A review on job scheduling techniques in cloud computing. *Journal of King Saud University – Computer and Information Sciences, 34*(7), 4390–4407. https://doi.org/10.1016/j.jksuci.2022.03.027

Omotunde, A., & Okolie, S. (2020). Resource allocation in cloud computing: An exposé. *Journal of Network and Computer Applications, 156*, 102577. https://doi.org/10.1016/j.jnca.2020.102577

Pradhan, P., Behera, P., & Ray, B. (2021). Resource allocation methodologies in cloud computing. In *Cloud Computing Technologies and Applications* (pp. 125–144). CRC Press. https://doi.org/10.1201/9781003337218-6

Sanjay, N. (2023). Optimized task scheduling and VM allocation in cloud computing. *Informatica, 47*(6), 1021–1032. https://doi.org/10.31449/inf.v47i6.7970

Saravanan, G., Neelakandan, S., Ezhumalai, P., & Maurya, S. (2023). Wild horse optimization with Levy flight algorithm for cloud task scheduling. *Journal of Cloud Computing, 12*(1), 1–18. https://doi.org/10.1186/s13677-023-00401-1

Shafiq, D., Jhanjhi, N., Abdullah, A., & Alzain, M. (2021). A load balancing algorithm for cloud data centers. *IEEE Access, 9*, 72976–72985. https://doi.org/10.1109/ACCESS.2021.3065308

Yadav, M., & Mishra, A. (2023). Enhanced ordinal optimization for task scheduling in cloud computing. *Journal of Cloud Computing, 12*(1), 45–60. https://doi.org/10.1186/s13677-023-00392-z