Libyan Journal of Educational Research and E-Learning (LJERE)
ISSN: 3105-3556
Volume 1, Issue 2, 2025
Page No: 290-307
Website: https://ljere.com.ly/index.php/ljere/index

Resource Allocation Challenges and Solutions in Cloud Computing for Machine Learning Workloads: A Comprehensive Review

Khalid Ahmed Masoud Qadoua [1*], Elbashir Mohamed Abdullah Khalil [2]
[1,2] Computer Science Department, Faculty of Science, Gharyan University, Libya

تخصيص الموارد في الحوسبة السحابية لتشغيل نماذج تعلّم الآلة التحديات والحلول مراجعة شاملة

خالد أحمد مسعود قدوع [1*]، البشير محمد عبدالله خليل [2]
[2،1] قسم علوم الحاسوب- كلية العلوم- جامعة غريان، ليبيا

*Corresponding author: kadouakhaled@gmail.com

Received: August 29, 2025	Accepted: November 15, 2025	Published: December 02, 2025

Abstract

In the rapidly evolving landscape of digital computation, the integration of cloud and fog computing stands at the forefront, offering unprecedented scalability and flexibility. This review paper delves into the critical realm of resource allocation within cloud computing, exploring various techniques and their implications, especially in the context of burgeoning machine learning applications. While cloud environments are revolutionizing how data-driven tasks are approached and executed, they concurrently pose intricate challenges, especially concerning efficient resource distribution. By examining related works and methodologies, this review provides a comprehensive understanding of current solutions and the inherent challenges they aim to address. As machine learning tasks become increasingly prevalent within the cloud, the nuances of resource allocation become even more pronounced, demanding innovative solutions. This paper encapsulates these nuances, charting a path for future research and highlighting the immense potential waiting to be unlocked in the confluence of cloud computing and machine learning.

Keywords: Resource Allocation, Cloud Computing, Machine Learning Fog Computing, Efficient Distribution, Technological Advancement, Scalability, Flexibility.

الملخص

في ظل التطور المتسارع في عالم الحوسبة الرقمية تبرز تكاملية الحوسبة السحابية والضبابية في الصدارة لما توفره من قدرات غير مسبوقة من حيث قابلية التوسع والمرونة ، يتناول هذا البحث بالاستعراض والتحليل أحد أهم المحاور في الحوسبة السحابية وهو تخصيص الموارد من خلال استكشاف أبرز التقنيات المستخدمة خصوصًا في سياق التطبيقات المتنامية لتعلم الآلة وعلى الرغم من الدور الريادي للبيئات السحابية في إعادة تشكيل طريقة تنفيذ المهام المعتمدة على البيانات فإنها تطرح في الوقت ذاته تحديات معقدة لا سيما فيما يتعلق بالتوزيع الفعّال للموارد من خلال مراجعة الأعمال البحثية ذات الصلة والمنهجيات المطروحة ويقدم هذا الاستعراض فهمًا شاملًا للحلول الحالية والتحديات الجوهرية التي تسعى هذه الحلول

لمعالجته ومع ازدياد انتشار مهام تعلم الآلة في البيئات السحابية تصبح دقة وفعالية تخصيص الموارد أكثر أهمية مما يستدعي حلولًا مبتكرة.
يلخص هذا البحث تلك الجوانب الدقيقة ويرسم مسارًا للبحوث المستقبلية مسلطًا الضوء على الإمكانات الكبيرة التي يمكن إطلاقها عند تقاطع الحوسبة السحابية وتعلم الآلة.

الكلمات المفتاحية : تخصيص الموارد، الحوسبة السحابية، تعلم الآلة، الحوسبة الضبابية، التوزيع الفعّال، التقدم التكنولوجي، قابلية التوسع، المرونة

Introduction

Fog computing stands as an emerging paradigm in distributed computing, offering a novel model that furnishes storage, communication, and computational capabilities in close proximity to end-users. This innovative concept represents an extension of the cloud computing framework, originally introduced by CISCO. Its core purpose is to address the escalating demands of internet users by enabling the processing of data in the vicinity of Internet of Things (IoT) devices, obviating the need to transmit data to a remote cloud infrastructure [1].

The impetus behind the development of fog computing stems from the mounting challenges associated with centralized cloud solutions. The manifold array of services that users request, coupled with the anticipation of real-time responses and rapid data exchanges, often overwhelms traditional cloud-based systems. Furthermore, many network edge devices remain underutilized in terms of their computational and storage capacities. In response to these quandaries, fog computing has emerged as a viable solution. It serves to furnish localized services to mobile users by assuming an intermediary role between the cloud and IoT devices [2].

Across the telecommunications landscape, numerous network operators have embarked on the provision of storage, computation, and communication facilities at the network's edge, effectively crafting an environment conducive to fog computing. This strategic shift allows bandwidth-intensive and real-time applications to be processed with enhanced cost efficiency and minimized latency [3].

Resource allocation within the context of a fog network involves the strategic selection of efficient resources, encompassing both cloud-based resources and those present within nearby fog nodes, to cater to the requests originating from IoT users. This approach serves to enhance responsiveness and meet Quality of Service (QoS) requirements for these requests [4]. Such resource allocation can be carried out in either a static or dynamic manner in the fog environment. However, it's important to note that static resource allocation can lead to high operational costs for IoT services. Allocating excessive resources to ensure QoS compliance can result in resource underutilization and increased costs. Conversely, insufficient resource allocation may lead to overutilization and compromised QoS standards. Consequently, achieving efficient and effective resource allocation becomes imperative to address the pitfalls of both over and under-provisioning [4].

The challenges encountered in resource allocation differ significantly between cloud and fog computing. In the realm of fog computing, the allocation of heterogeneous and unpredictable fog nodes to execute diverse service requests with varying QoS requirements poses a distinct set of challenges. Fog computing comprises a multitude of entities including IoT users, fog nodes, and cloud servers. Consequently, resource allocation in this fog environment is just as intricate as in cloud computing and cannot be tackled with existing resource allocation techniques alone. This complexity arises from limitations in resources, their heterogeneity, and the dynamic and uncertain nature of the fog environment [5].

The primary goal of resource allocation in fog computing is to optimally assign the best available resources to tasks generated by edge devices, ensuring QoS requirements are met. Ongoing research in this domain highlights the ongoing evolution of resource allocation

techniques. However, selecting the most efficient and appropriate resource allocation algorithm remains a challenge [5]. Conventional resource allocation methods often fall short in addressing the unique demands of fog environments. This has spurred the exploration of various approaches rooted in heuristic and metaheuristic methods to achieve optimal solutions. Heuristic methods, while effective, can sometimes become trapped in local minima issues. In contrast, metaheuristic approaches are more efficient and adept at avoiding local minima problems [6].

Metaheuristic approaches have demonstrated their efficacy in comparison to heuristic methods, showcasing improved results in terms of QoS and computational efficiency. These approaches can be applied to a wide array of real-life optimization problems to enhance overall efficiency and performance. Notably, the application of metaheuristic approaches in resource allocation for fog computing remains largely unexplored. This paper's central objective is to provide a comprehensive overview of available metaheuristic approaches for resource allocation within fog computing. Furthermore, it aims to outline potential avenues of future research in this rapidly evolving field. The main contributions of this review paper are as follows:

- We propose a clear and structured taxonomy of resource allocation approaches for ML workloads across cloud and fog environments, covering heuristic, optimization-based, machine-learning, and hybrid methods.
- We provide a comparative examination of cloud vs. fog resource allocation strategies, highlighting their trade-offs in latency, scalability, energy efficiency, and suitability for ML training and inference.
- We assess existing approaches using unified performance metrics such as latency, throughput, accuracy, cost efficiency, and resource utilization, revealing the strengths and limitations of each class.
- We identify key research gaps—including real-time ML scheduling, unified cloud–fog orchestration, and GPU/TPU allocation and outline promising directions for next-generation intelligent resource management.

The remainder structure of this paper is systematically arranged to ensure clarity an comprehensiveness. Section 2 goes into detail on the concepts and background of cloud and fog computing that are described. Then, Section 3 describes Resource Allocation in Cloud and Fog computing. Section 4 provides resource allocation techniques, Section5 presents Related work in RL and DLR Finally, Section6 wraps up our discussion with a conclusion, summarizing the key takeaways from our study.

Cloud and Fog Computing

Clouds and fog are manifestations of the atmosphere's ability to condense water vapor into tiny liquid droplets or, under certain conditions, into ice crystals [7]. The primary difference between them lies in their spatial occurrence. Clouds are typically found at various altitudes in the troposphere, from the near-surface boundary layer to elevations exceeding several kilometers [8]. Their formation is a result of air rising and cooling to its dew point or by the addition of moisture. The resulting condensation forms around aerosols, minute particulate matter like dust, salt, or soot, which act as condensation nuclei. Depending on the altitude, temperature, and atmospheric conditions, clouds can take on diverse morphologies, from the puffy cumulus clouds often associated with fair weather to the high-altitude, wispy cirrus clouds composed mainly of ice crystals. Conversely, fog is a low-lying cloud, occurring when the immediate surface air reaches its saturation point, leading to condensation [9]. This is frequently a result of radiational cooling during calm, clear nights where the Earth's surface loses heat rapidly, or from warm air moving over a colder surface, such as a body of water. Fog, given its ground-level existence, has significant implications for human activities,

especially transportation, as it can drastically reduce visibility, necessitating the need for added caution in navigation, both on roads and at sea.

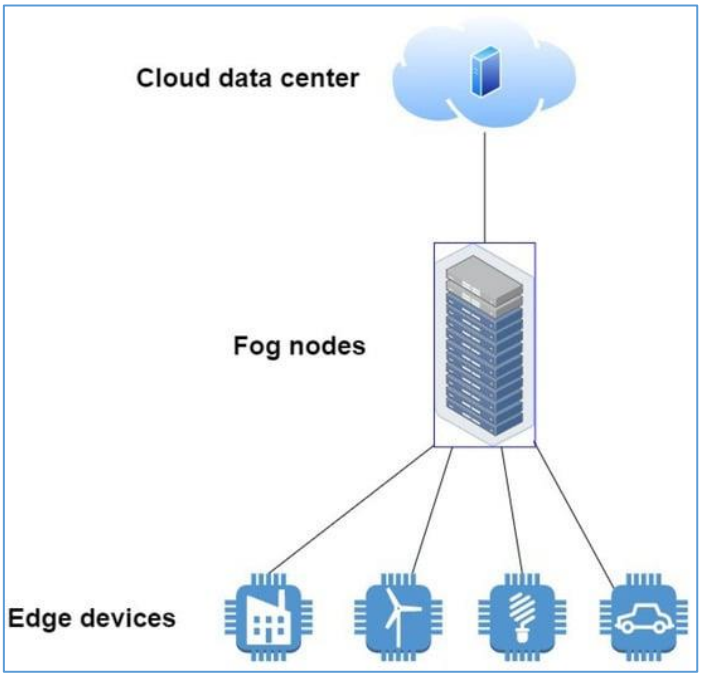

Figure 1: *Cloud, fog and edge interconnection [10].*

Resource Allocation in Cloud and Fog computing

Efficient resource allocation is a critical challenge in both cloud and fog computing environments due to the heterogeneous nature of resources, dynamic workloads, and diverse application requirements. Cloud computing offers centralized, large-scale resource pools with elastic scalability, while fog computing extends these capabilities closer to end-users, enabling low-latency services and context-aware processing [9].

Resource allocation in the realm of cloud computing involves a systematic procedure through which virtual machines are assigned with the aim of meeting the specifications outlined by consumers. This process extends to determining the most effective manner in which workloads can be distributed among virtual machines, subsequently managing them in an optimal fashion within the cloud environment [11]. Essentially, it revolves around establishing the commencement and conclusion of computational actions based on several factors, including: 1) the allocation of resources, 2) the time required for execution, 3) the actions of preceding tasks, and 4) the interdependencies among these preceding tasks. This process of resource allocation is visualized in Figure 2, depicting the overall sequence of steps and interactions involved in the allocation of resources.

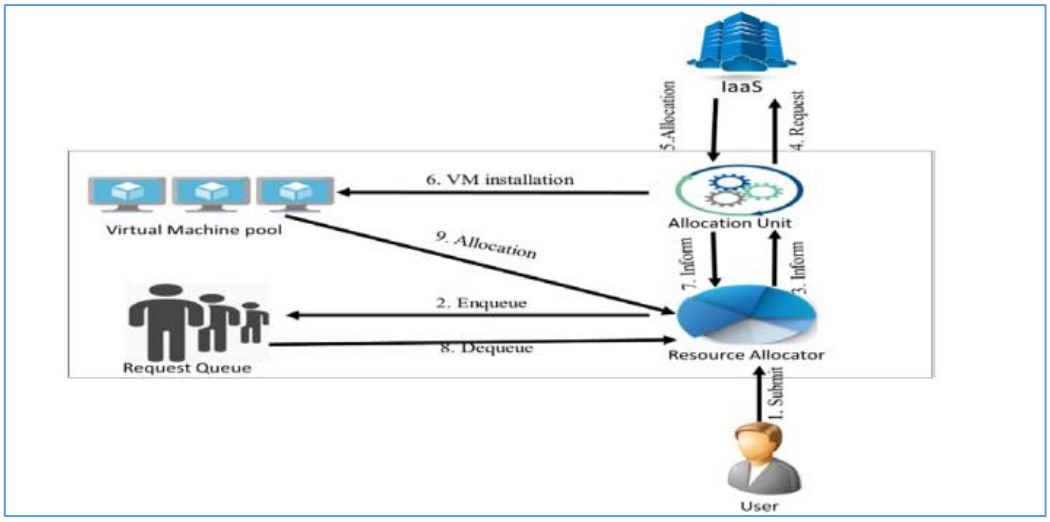

Figure 2: *resource allocation in CC.*

Furthermore, within the data centers of cloud service providers, resources exhibit an escalating level of heterogeneity, encompassing a range of evolving generations of hardware infrastructure driven by technological advancements. In contemporary industries, multicore CPUs augmented with substantial cache memory have become commonplace. Notably, data centers now feature CPUs designed for energy efficiency, such as the RM CPU [12] and Atom CPU [13], poised for deployment. Beyond processors, other subsystems are also undergoing rapid evolution. These advancements encompass novel technologies like Memristors [14], Phase Change Random Access Memory [15], as well as storage and memory subsystems utilizing Solid State Drives [16]. The very architecture of networks is also in flux, exemplified by innovations like B-Cube [17] and D-Cell [18].

To remain at the forefront and ensure their services remain current, cloud service providers must readily integrate these technological shifts. However, embracing these advancements accentuates the challenge of managing this growing heterogeneity. Therefore, it becomes imperative to harness this ever-expanding heterogeneity as a means for both cloud service providers and consumers to achieve their respective goals: optimizing resource utilization while maintaining cost efficiency.

Conversely, cloud service providers encounter a range of formidable challenges when allocating resources among users' tasks based on their application usage patterns. A subset of these challenges is outlined below:

1. Anticipating the specific application requirements of consumers proves to be a complex endeavor for cloud service providers. Meanwhile, consumers expect their tasks to be accomplished within stipulated timeframes. Hence, the development of efficient resource allocation techniques becomes imperative to surmount this predicament.
2. The physical machines within the cloud infrastructure must possess the capacity to adequately cater to the resource demands of each virtual machine operating on them. Concurrently, consumers demand networking services characterized by effective Quality of Service (QoS) to ensure the seamless transmission of their application data.
3. For tasks that may exceed standard completion durations, service providers must strategically schedule resource availability. This underscores the need for a technique capable of managing interruptions and seamlessly transitioning tasks to available resources.
4. The pursuit of energy-efficient resource allocation poses a paramount challenge within the realm of cloud computing. Escalating energy costs and the imperative to curtail greenhouse

gas emissions have prompted a drive to minimize overall energy consumption, encompassing communication, and storage.

The discourse above underscores the intrinsic importance of considering the attributes and characteristics of both cloud service providers and consumers. This holistic approach is essential to furnish efficient services, ensuring that suitable tasks receive adequate resource allocations, thereby facilitating timely task completion and enabling cloud service providers to optimize their profits.

Resource Allocation Techniques

The domain of resource allocation within cloud computing necessitates strategic decisions on the part of the cloud service provider, encompassing determinations regarding what, when, how much, and where to allocate the available resources for tasks. Typically, users specify the quantity and nature of resources required for their requests, prompting service providers to assign the requested resources within their data centers. To ensure effective application execution, the resource container types and quantities must align with user-defined constraints, such as job completion time deadlines, and exhibit compatibility with the workload characteristics.

This study's analysis reveals that resource allocation techniques can be classified into five distinct categories:

1. **Strategic:** Adapting to the dynamic demands of consumers.
2. **Target Resources:** Focusing primarily on fulfilling requested resources.
3. **Optimization:** Maximizing resource utilization through optimization.
4. **Scheduling:** Prioritizing tasks for enhanced performance.
5. **Power:** Achieving improved resource allocation with reduced power consumption. A schematic overview of this taxonomy is presented in Figure 3

A range of parameters exist to evaluate diverse resource allocation techniques from both the cloud service provider and consumer perspectives:

Cloud Service Provider:

- **Resource Utilization:** Ensuring efficient use of all available resources to avert idleness and contribute to environmental sustainability and profit maximization.
- **Workload:** Verifying that the system workload is sufficient to meet task completion deadlines.
- **Cost:** Gauging the financial ramifications for the cloud service provider, while considering profit and loss.
- **Energy:** Minimizing energy consumption to align with environmental sustainability goals.
- **SLA (Service Level Agreement) and QoS (Quality of Service):** Ensuring contractual commitments and high-quality service provision.

Cloud Service Consumer:

- **Response Time:** Minimizing the system's response time to enhance overall performance.
- **User Satisfaction**: Maximizing user satisfaction through effective resource allocation.
- **Execution Time:** Striving for minimal execution time for both provider and consumer benefits.
- **QoS and SLA:** Ensuring service quality and adherence to service agreements.

These parameters collectively provide a comprehensive framework to evaluate and compare various resource allocation techniques within the cloud computing landscape.

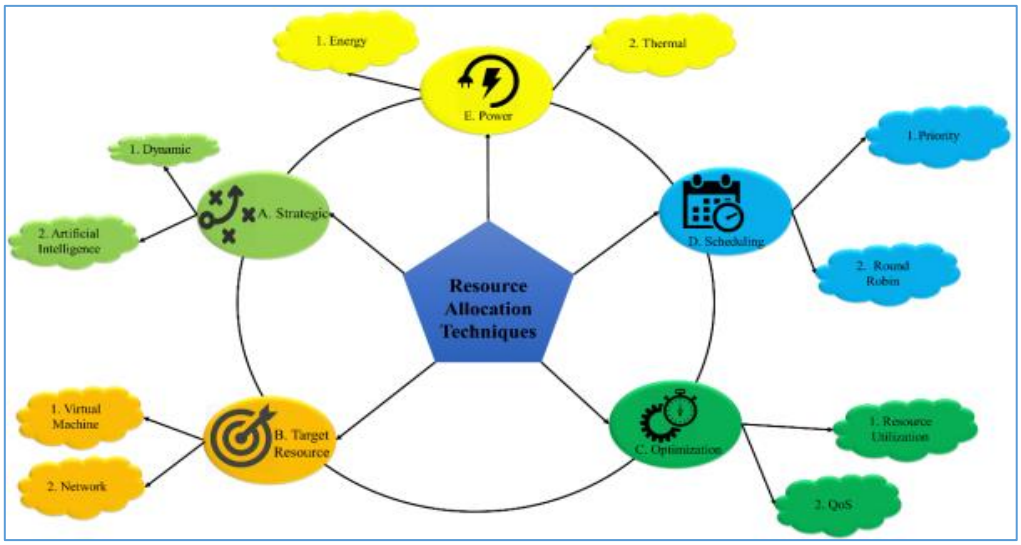

Figure 3: *Taxonomy of resource allocation techniques.*

Related work

Resource allocation in emerging network technologies has witnessed considerable interest, leveraging the capabilities of Reinforcement Learning (RL). One notable approach is encapsulated in Ref. [19], which proposes a radio resource allocation model tailored for 5G haptic communications. Using Q-learning, it seeks to optimize the use of scarce radio resources, taking into account the dynamic nature of vertical haptic applications. Their system uniquely characterizes the state space using factors such as allocated resources and application performance requirements. Evaluation of their approach utilizes a seven-cell hexagonal grid model, distinguishing between radio slices for haptic and human-to-human communications.

In a similar vein, Ref. [20] advances a resource allocation framework designed for 5G Fog-RAN, using a range of RL algorithms like Q-learning, SARSA, expected SARSA, and Monte Carlo. The framework aims to fulfill the stringent low-latency needs of IoT applications while promoting resource efficiency. The state space is defined by resource blocks and IoT application attributes. Their evaluation landscape involves ten varied IoT applications, ranging from smart farming to connected health.

Diving deeper into 5G Fog-RAN, Ref. [21] presents a Q-learning-based method focusing on reducing latency and optimizing computing resource utilization. The system's state space factors in user requests and resource allocation metrics. The evaluation is grounded on an open-source 5G K network simulator, employing openAI gym for Q-learning implementation. Meanwhile, the realm of low-orbit satellite networks is explored by Ref. [22], which suggests a dynamic resource allocation strategy to maximize provider revenue and enhance users' Quality of Service (QoS). This approach, also based on Q-learning, uses satellite slices emulated through Mininet.

Several other endeavors, like Refs. [24,25], target 5G-RAN, particularly emphasizing eMBB and vehicle-to-everything (V2X) services. The proposed Q-learning scheme focuses on maximizing resource use, considering the bandwidth's resource blocks. On a related note, Ref. [26] integrates both Monte Carlo and Q-learning to optimize power resource allocation in an edge-computing environment, while Ref. [27] pursues Q-learning for improving various metrics, including latency, in 5G Fog-RAN.

Transitioning from Q-learning, Ref. [23] employs PPO (proximal policy optimization) for dynamic resource allocation, aiming for enhanced resource efficiency in multilayer mobile

edge computing domains. Their approach uniquely weighs service type and resource utilization metrics. Similarly, Ref. [28] introduces a two-tier Q-learning strategy, designed to elevate operators' revenue in a multitenant 5G network environment. The methodology entails a dual-phase approach: VNF mapping followed by user association and power allocation.

Table 1: *Related work in RL*

Ref.	Approach	Algorithm	Network Focus	Key Features	Evaluation Tool
19	Radio resource allocation for 5G haptic communications	Q-learning	5G haptic communications	Optimizes scarce radio resources, considers application performance requirements	Seven-cell hexagonal grid model
20	Resource allocation framework for 5G Fog-RAN	Q-learning, SARSA, expected SARSA, Monte Carlo	5G Fog-RAN	Low-latency for IoT, state space defined by resource blocks and IoT attributes	IoT environment with ten applications
21	Resource allocation method for 5G Fog-RAN	Q-learning	5G Fog-RAN	Reduces latency, optimizes computing resources	Open-source 5G K network simulator, openAI gym
22	Dynamic resource allocation for low-orbit satellite networks	Q-learning	Low-orbit satellite networks	Maximizes provider revenue, enhances user QoS	Mininet (emulating satellite slices)
24,25	Resource allocation scheme for 5G-RAN	Q-learning	5G-RAN	Focus on eMBB and V2X services, bandwidth's resource blocks	Not specified
26	Power resource allocation for edge-computing environment	Monte Carlo, Q-learning	Edge-computing	Binary space of states, power level allocation	Not specified
27	Resource allocation for 5G Fog-RAN	Q-learning	5G Fog-RAN	Improves latency, energy consumption, and cost metrics	Not specified
23	Dynamic resource allocation for mobile edge computing	PPO	Multilayer mobile edge computing	Service type and resource utilization metrics	Not specified
28	Two-tier resource allocation strategy for 5G	Two-stage Q-learning	Multitenant 5G	VNF mapping, user association, and power allocation	Not specified

Critical Discussion and Identification of Research Gaps in RL

In Figure 4 provides a visual representation of the research gaps found in papers focused on resource allocation within the realms of RL and Deep RL. In this layout, each row corresponds to a specific paper, while the columns indicate various gap categories. If you see a "1" (shown as a dark cell), it signals that a gap is present; conversely, a "0" (represented as a light cell) means there's no gap. The heatmap makes it clear that many of these papers struggle with a limited set of evaluation metrics, often zeroing in solely on aspects like Quality of Service (QoS), latency, or throughput. What's particularly striking is the widespread lack of evaluation on system-level efficiency—things like CPU usage, memory, and bandwidth are often overlooked. This points to a broader issue: the absence of a comprehensive performance analysis. Moreover, multi-objective optimization seems to be a rather neglected area. Most studies tend to focus on optimizing just one metric, without considering the necessary trade-offs that come into play. Lastly, many papers are constrained by their focus on specific architectures or application domains, which ultimately restricts their generalizability. In summary, the figure clearly highlights that these gaps are not just random occurrences; they reflect a systematic issue within the field. This calls for more holistic, adaptable, and multi-metric strategies for resource allocation.

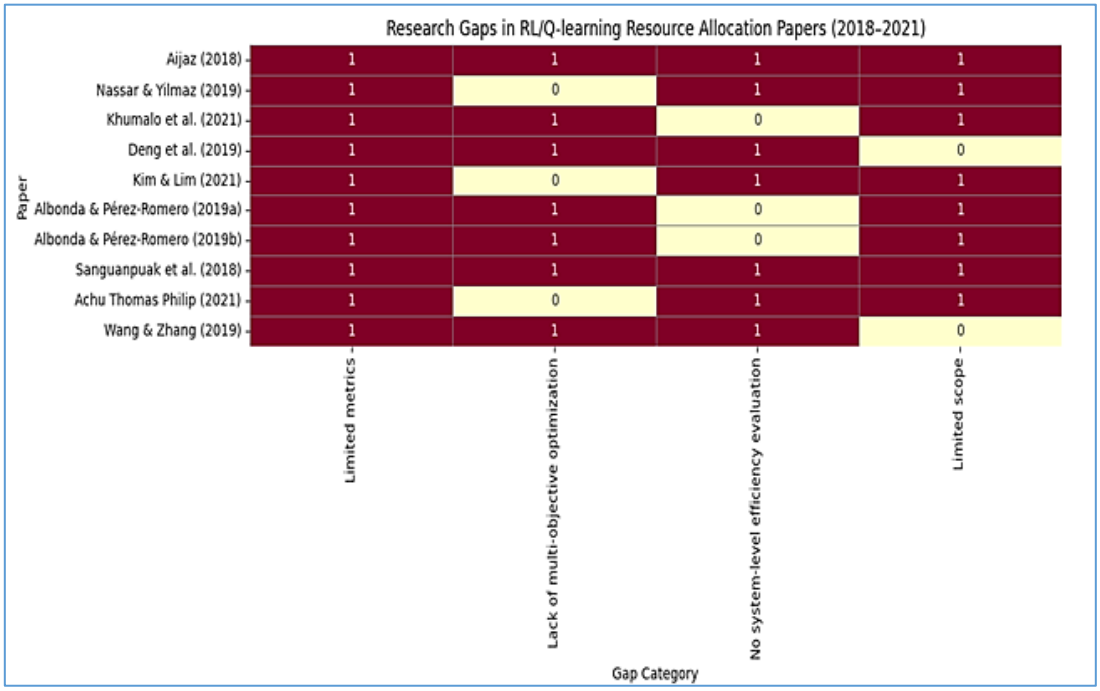

Figure 4:*Heatmap of Research Gaps Across RL/Deep RL Resource Allocation Papers (2018–2022).*

Recent studies have delved deeply into the application of deep reinforcement learning (DRL) for the optimization of resource allocation in network slice lifecycle (NSL) and related technologies. In one notable study, a framework was put forth that targets the allocation of computing, storage, and radio resources in manufacturing services with the primary aim of maximizing the long-term incomes for providers [29,30]. To accomplish this, they deployed Double Deep Q-Networks (DDQN) and dueling DQN agents. These agents were fine-tuned using the stochastic gradient descent (SGD) algorithm and constructed their states grounded on requested resources alongside the available computing resources. The connectivity capabilities of data centers, which are designated for storing virtualized radio resources, were also taken

into account. To evaluate the model, an emulated environment was designed in TensorFlow, incorporating three slice classes: utilities, automotive, and manufacturing. Importantly, slice requests in this setup adhered to a Poisson distribution.

Another study presented a strategy rooted in DQN, aiming to allocate both radio and backhaul resources within a virtualized Radio Access Network (RAN) [31]. The purpose was to strike a balance between the Quality of Service (QoS) satisfaction and resource utilization for slices. Meanwhile, a distinct proposal utilized a custom-designed RL algorithm in conjunction with a Deep Neural Network (DNN) to hone the positioning of functions that comprise service function chains specifically in metro-core optical networks [32]. The DNN comprised four hidden layers, each with a count of 100 nodes. The method incorporated a state space that considered three layers: optical, IP/MPLS, and service slicing. This proposal's evaluation framework, along with its algorithm, was established in the openAI gym environment, and a mobile traffic dataset from the Milan urban area was employed for training [33].

Shifting focus to 5G-RAN, one framework was introduced which leaned on a constrained DQN developed through a DNN [37]. The structure of this DNN consisted of two connected layers having 64 and 32 nodes respectively. This framework primarily sought to meet the performance requisites of distinct slices, including video, VoLTE, and uRLLC. On a related note, there was an innovative approach that incorporated a powered DDQN to address both service level agreements (SLAs) and enhance the efficiency of resource utilization [35]. This was all in light of the service request dynamics on 5G-RAN.

Delving deeper into 5G-RAN, an approach was elucidated that judiciously allocated resources aiming to elevate the satisfaction level for network slice requirements [39,40,41]. This method visualized the space of states as the available radio resource blocks. The actions, on the other hand, were conceptualized as the allocation of resource blocks or their non-allocation. As a technique to quicken learning, the Ape-X method was integrated, facilitating the concurrent processing by multiple DQN agents.

In yet another intriguing study, a two-tier approach for resource allocation was brought to the fore [42]. This strategy aimed at addressing the Quality of Experience (QoE) requisites and attaining efficient utilization on 5G end-to-end slicing. A different mechanism, reliant on DQN, endeavored to allocate bandwidth alongside 5G-RAN resources to serve a variety of applications, ranging from mobile and videos to vehicle communications [43]. The primary aspiration here was to escalate the long-term resource utilization and subsequently the revenues for virtual network providers.

Further, in the realm of bandwidth and virtual machine allocation, a DQN-based methodology was propounded that catered to services either queued within a time frame or during their arrival [44]. Another work employing both DQN and DDQN centered on elastic and real-time slices, and the focal point here was to augment spectral efficiency utilization in 5G-RAN, especially when numerous intelligent devices are in play [45].

Several other studies have also made significant contributions in this domain. For instance, a dynamic framework using DQN aimed to reserve and allocate unused bandwidth resources in virtualized RAN [46,47]. Another DQN-based approach was crafted to manage resources, specifically targeting cache reservation and allocation at edge networks [48]. Unique bandwidth allocation strategies employing Long Short-Term Memory (LSTM)-based advantage actor–critic (A2C) algorithms were also highlighted, aiming to optimize spectral efficiency and SLA satisfaction ratios [36]. Lastly, there have been works which specifically targeted Internet vehicular and smart city applications using DQN [34], as well as those that employed multi-agent DQNs in conjunction with an SDN controller for optimal radio resource allocation [50].

Table 2: *Relatedwork in DLR*

References	Algorithm	Target	Key Features	Data
[29,30]	DDQN and dueling DQN agents	Maximize long-term incomes for providers	States: Requested resources, available computing resources. Actions: Resources to assign per request	TensorFlow with three slice classes
[31]	DQN-based strategy	Balance QoS satisfaction & resource utilization	States: Satisfaction ratio, resources allocated. Actions: Optimal resource provisioning	Emulated mobile network with four slice classes
[32]	Proprietary RL algorithm & DNN	Optimize positioning of functions in metro-core optical networks	States: Optical, IP/MPLS, and service slicing layers. Actions: Reconfiguration decision	openAI gym with Milan urban area mobile traffic dataset
[37]	Constrained DQN with DNN	Meet requirements of video, VoLTE, uRLLC slices	States: Active users per service. Actions: Bandwidth allocation	5G-RAN setting
[35]	Powered DDQN with GANs	Meet SLAs, maximize resource utilization & provider revenue	States: Service demands within time. Actions: Bandwidth assignment	5G-RAN dynamics
[39,40,41]	DQN-based approach	Maximize slice requirement satisfaction & resource block usage	States: Available radio resource blocks. Actions: Allocation decisions	Utilized Ape-X method for learning acceleration
[42]	Two-tier DQN approach	Meet QoE requirements & efficient utilization	States: Available radio units, QoE satisfaction. Actions: Access unit formations	5G end-to-end slicing
[43]	DQN-based mechanism	Increase long-term resource utilization & provider revenue	States: Requested bandwidth, consumed energy. Actions: Slice selection	Mobile, videos, vehicle communications in 5G-RAN
[44]	DQN-based approach	Optimize delays & resource usage costs	States: Resource request arrivals, queue levels. Actions: Bandwidth allocation	Time window or arrival-based service queueing
[45]	DQN & DDQN approach	Maximize spectral efficiency & reduce costs	States: Carrier power traffic per slice. Actions: Bandwidth allocation	5G-RAN with multiple intelligent devices
[46,47]	DQN-based framework	Maximize QoS satisfaction &	States: Allocated virtual resources, average	Virtualized RAN setting

		resource utilization	resource utilization. Actions: Resource adjustment	
[48]	DQN-based approach	Maximize QoS satisfaction & network utilization	States: Resource utilization, reserved resources. Actions: Cache resource allocation	Edge network with mobile virtual network operators
[36]	LSTM-based A2C	Maximize spectral efficiency, SLA satisfaction & profit	States: Slice arrival requests. Actions: Bandwidth allocation	RAN setting
[34]	DQN-based solution	Allocate resources for vehicular & smart city apps	States: Resource blocks at time t. Actions: User request decisions	Cloudified RAN or edge network
[49]	Soft actor–critic algorithm	Maximize throughput	States: Channel, battery, queue state. Actions: Subchannel allocation, harvesting time	Discrete channel & continuous energy-harvesting division
[50]	Multiagents DQN with SDN controller	Maximize data rate	States: Set of end-users, channel gain, data rate, delay. Actions: Resource block assignments	Radio resources for uRLLC&eMBB

In recent developments within the domain of network function virtualization (NFV) [51], there is a growing emphasis on substituting physical middleboxes with the more adaptive virtual network functions (VNFs). One of the core challenges addressed in this transformation is the dynamic adaptability to the incessantly fluctuating traffic demands. While VNFs offer the advantage of instantiation on the fly, adjusting their allocated resources based on real-time needs is a sophisticated decision-making process.

Traditional optimization methodologies often operate on the presumption of consistent resource requirements for every individual VNF instance. This, however, presents a pitfall; setting resources that are either too constrained or too expansive can result in inefficient resource utilization or compromised service quality.

Addressing this conundrum, there has been an approach that harnesses machine learning to predict VNF resource demands. By training models on authentic VNF datasets, which encapsulate performance metrics and resource needs, the tailored models can then forecast the resource prerequisites for managing specific traffic loads. Integrating these machine learning models into joint algorithms for VNF scaling and placement has furthered the potential for efficient resource allocation. An empirical assessment grounded in real-world data vouched for the efficacy of this approach. Notably, the utilization of these optimized machine learning models led to a significant curtailment in resource wastage. The findings revealed a staggering reduction in resource consumption by up to 12-fold. In tandem, there was a marked improvement in service quality, with delays reduced to a fraction, up to 4.5 times less than what's observed with conventional fixed resource allocations.

Critical Discussion and Identification of Research Gaps in Deep RL

As we illustrates in figure 5 offers a straightforward look at how research gaps are spread out across papers that delve into resource allocation in RL and Deep RL (29–51). If you take a peek at the heatmap, it's pretty clear that many studies face a common issue—they often rely on a narrow set of evaluation metrics. Most of them seem to focus solely on aspects like Quality of Service (QoS), latency, or throughput, while they overlook more comprehensive operational measures like CPU usage, memory efficiency, and how much bandwidth is being consumed. What's striking is the widespread lack of evaluation when it comes to system-level efficiency. This suggests that many of these approaches haven't really been tested under real-world resource constraints. And then there's the noticeable gap in multi-objective optimization. It's like most of the previous research has zeroed in on optimizing just one performance metric, without considering the trade-offs between latency, energy consumption, cost, and QoS.Finally, some papers exhibit scope limitations, being tied to specific architectures or application domains (e.g., smart grid, vehicular systems, NFV), which reduces the generalizability of their findings. Overall, the heatmap highlights that these gaps are systematic rather than isolated, underscoring the need for a comprehensive, flexible, and multi-metric framework for resource allocation.

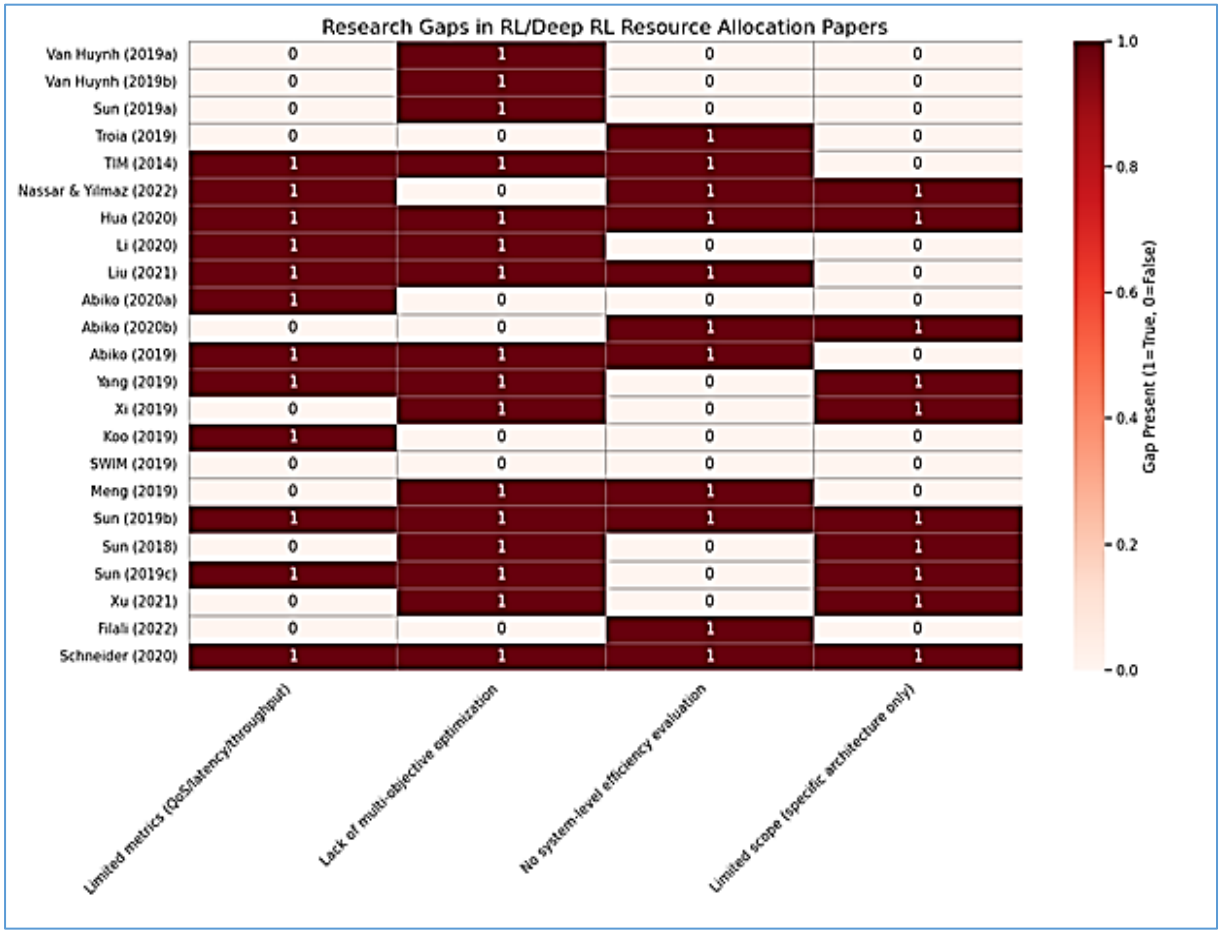

Figure 5:*Heatmap of Research Gaps Across RL/Deep RL Resource Allocation Papers (29–51)*

Challenges of Resource Allocation in Machine Learning in Cloud and fog Environments

Resource allocation for machine learning in cloud and fog environments presents its own unique set of challenges, a testament to the intricate intersection of two evolving fields. First and foremost, the dynamic nature of machine learning workloads, characterized by a mix of

training, inference, and data preprocessing tasks, demands a high degree of flexibility in allocating resources. Unlike traditional applications that might have predictable usage patterns, ML workflows can vary significantly in their resource needs depending on the phase, model complexity, and dataset size.

Additionally, the large-scale datasets inherent to many machine learning tasks put strain on storage and data transfer resources in cloud environments. This means that not only compute resources, but also data pipelines and storage infrastructures, must be efficiently managed to prevent bottlenecks that can drastically slow down model training and inference. Ensuring seamless data transfer between distributed data sources and processing nodes while minimizing latency is a Herculean task in its own right.

The heterogeneity of cloud resources further complicates the issue. Cloud environments typically consist of a mix of CPUs, GPUs, TPUs, and other specialized hardware accelerators. Each of these has its own strengths and trade-offs, and machine learning tasks can show varying levels of performance depending on the hardware they are run on. Efficiently mapping tasks to the most appropriate hardware, while maximizing utilization and minimizing costs, requires sophisticated scheduling and orchestration mechanisms.

Moreover, ensuring fault tolerance and scalability becomes vital, especially when training large models or serving popular ML applications. Machine learning processes, especially training, can be prolonged and consume significant computational resources. Ensuring that these processes are resilient to failures, can be resumed in case of interruptions, and can be scaled out to leverage more resources as needed, becomes imperative.

Lastly, the multi-tenancy nature of cloud environments, where resources are shared among multiple users and tasks, poses challenges in guaranteeing fairness and isolation. Ensuring that one user's resource-intensive machine learning task does not starve other tasks of necessary resources, while still meeting quality-of-service guarantees, requires a delicate balancing act.

In essence, while the fusion of machine learning with cloud environments offers immense potential, it comes with its own set of intricate challenges that require both deep technical expertise and innovative solutions to effectively tackle.

Conclusion

In the realm of modern computing paradigms, the synergy of cloud and fog computing has undeniably ushered in a new era of scalability, efficiency, and flexibility. This review delved deep into the intricacies of resource allocation within cloud computing, shedding light on the variety of techniques and methodologies currently in practice. These allocation strategies, vital to the efficient operation of cloud environments, are perpetually evolving to meet the dynamic and diverse demands of applications and services, especially with the surge in machine learning tasks. Through our examination of related works, it's evident that the academic and industry landscapes are rife with pioneering solutions and innovations, yet they grapple with inherent challenges.

As we highlighted in our discourse on the challenges of resource allocation specific to machine learning in cloud settings, the confluence of these two powerful domains, while promising, is fraught with complexities. The dynamic nature of ML workloads, data-intensive demands, heterogeneity of resources, need for fault tolerance, scalability concerns, and the overarching principle of ensuring fairness in multi-tenancy environments all present intricate puzzles to be solved.

It's clear that the future of cloud computing and machine learning will necessitate a continuous and collaborative effort from researchers, practitioners, and industry stakeholders. The journey ahead, while challenging, also presents immense opportunities for breakthroughs that can reshape the contours of modern computing. As the boundaries of what's possible in the cloud

continue to expand, driven by the relentless pace of technological advancement, resource allocation will remain a focal point of interest, and its effective resolution will be pivotal in harnessing the full potential of the cloud.

References

1. Hu, P., Dhelim, S., Ning, H., & Qiu, T. (2017). Survey on fog computing: architecture, key technologies, applications and open issues. *J. Netw. Comput. Appl.*, vol. 98, pp. 27–42. [DOI](https://doi.org/10.1016/j.jnca.2017.08.010)
2. Chiang, M., & Zhang, T. (2016). Fog and IoT: An Overview of Research Opportunities. *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864. [DOI](https://doi.org/10.1109/JIOT.2016.2510075)
3. Mehta, A., Tarneberg, W., Klein, C., Tordsson, J., Kihl, M., &Elmroth, E. (2016). How Beneficial Are Intermediate Layer Data Centers in Mobile Edge Networks? In *FAS*W*, Augsburg, Germany, pp. 222–229.
4. El Kafhali, S., & Salah, K. (2017). Efficient and dynamic scaling of fog nodes for IoT devices. *J. Supercomput.*, vol. 73, no. 12, pp. 5261–5284. [DOI](https://doi.org/10.1007/s11227-017-2071-y)
5. Ghobaei-Arani, M., Souri, A., &Rahmanian, A. A. (2019). Resource Management Approaches in Fog Computing: a Comprehensive Review. *J. Grid Comput.*. [DOI](https://doi.org/10.1007/s10723-019-09488-9)
6. Tsai, C.-W., & Rodrigues, J. J. P. C. (2014). Metaheuristic Scheduling for Cloud: A Survey. *IEEE Syst. J.*, vol. 8, no. 1, pp. 279–291. [DOI](https://doi.org/10.1109/JSYST.2012.2227796)
7. Lamer, K., Oue, M., Battaglia, A., Roy, R. J., Cooper, K. B., Dhillon, R., & Kollias, P. (2021). Multifrequency radar observations of clouds and precipitation including the G-band. Atmospheric Measurement Techniques, 14(5), 3615-3629.
8. Davis, E. V., Rajeev, K., & Mishra, M. K. (2020). Effect of clouds on the diurnal evolution of the atmospheric boundary-layer height over a tropical coastal station. Boundary-Layer Meteorology, 175, 135-152.
9. Constantin, A. (2023). Exact nonlinear mountain waves propagating upwards. Journal of Physics A: Mathematical and Theoretical, 56(24), 245702.
10. Alwakeel, A. M. (2021). An overview of fog computing and edge computing security and privacy issues. Sensors, 21(24), 8226.
11. Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. NIST Special Publication. CrossRef Link
12. Selvi, S. T., Valliyammai, C., &Dhatchayani, V. N. (2014). Resource allocation issues and challenges in cloud computing. In Proc. of IEEE International Conference on Recent Trends in Information Technology, pp. 1-6. CrossRef Link
13. ARM—the architecture for the digital world. http://www.arm.com/. Accessed 18 January 2020.
14. Intel® AtomTM Processor. http://www.intel.com/content/www/us/en/processors/atom/atom-processor.html. Accessed 18 January 2020.
15. Memristor. http://www.memristor.org/.
16. Simpson, R. E., Fons, P., Kolobov, A. V., Fukaya, T., Krbal, M., Yagi, T., & Tominaga, J. (2011). Interfacial phase-change memory. Nature Nanotechnology, vol. 6, no. 8, pp. 501-505. CrossRef Link
17. Ekker, N., Coughlin, T., & Handy, J. (2009). Solid State Storage 101: An introduction to Solid State Storage. Storage Network Industry Association.

18. Guo, C., Lu, G., Li, D., Wu, H., Zhang, X., Shi, Y., ... Lu, S. (2009). BCube: a high performance, server-centric network architecture for modular data centers. ACM SIGCOMM Computer Communication Review, vol. 39, no. 4, pp. 63-74. CrossRef Link
19. Aijaz, A. (2018). Hap-SliceR: A Radio Resource Slicing Framework for 5G Networks with Haptic Communications. *IEEE Systems Journal*, 12, 2285–2296. (https://doi.org/10.1109/JSYST.2017.2719910)
20. Nassar, A., & Yilmaz, Y. (2019). Reinforcement Learning for Adaptive Resource Allocation in Fog RAN for IoT with Heterogeneous Latency Requirements. *IEEE Access*, 7, 128014–128025. (https://doi.org/10.1109/ACCESS.2019.2937909)
21. Khumalo, N. N., Oyerinde, O. O., &Mfupe, L. (2021). Reinforcement Learning-Based Resource Management Model for Fog Radio Access Network Architectures in 5G. *IEEE Access*, 9, 12706–12716. (https://doi.org/10.1109/ACCESS.2021.3057435)
22. Deng, Z., Du, Q., Li, N., & Zhang, Y. (2019). RL-Based Radio Resource Slicing Strategy for Software-Defined Satellite Networks. In *Proceedings of the 2019 IEEE 19th International Conference on Communication Technology*, Xi'an, China, 16–19 October 2019, pp. 897–901.
23. Kim, Y., & Lim, H. (2021). Multi-Agent Reinforcement Learning-Based Resource Management for End-to-End Network Slicing. *IEEE Access*, 9, 56178–56190. (https://doi.org/10.1109/ACCESS.2021.3078436)
24. Albonda, H. D. R., & Pérez-Romero, J. (2019a). Reinforcement Learning-Based Radio Access Network Slicing for a 5G System with Support for Cellular V2X. In *International Conference on Cognitive Radio Oriented Wireless Networks*, Springer, Cham, Switzerland, pp. 262–276.
25. Albonda, H. D. R., & Pérez-Romero, J. (2019b). An Efficient RAN Slicing Strategy for a Heterogeneous Network with eMBB and V2X Services. *IEEE Access*, 7, 44771–44782. (https://doi.org/10.1109/ACCESS.2019.2908701)
26. Sanguanpuak, T., Rajatheva, N., Niyato, D., &Latva-aho, M. (2018). Network Slicing with Mobile Edge Computing for Micro-Operator Networks in Beyond 5G. In *Proceedings of the 2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Chiang Rai, Thailand, 25–28 November 2018.
27. Achu Thomas Philip, N. M. (2021). Computation of 5G Fog-Radio Access Network Resource Allocation Scheme Using Reinforcement Learning. *International Research Journal of Engineering and Technology (IRJET)*, 8, 513–516.
28. Wang, X., & Zhang, T. (2019). Reinforcement Learning Based Resource Allocation for Network Slicing in 5G C-RAN. In *Proceedings of the 2019 Computing, Communications and IoT Applications (ComComAp)*, Shenzhen, China, 26–28 October 2019, pp. 106–111.
29. Van Huynh, N., Thai Hoang, D., Nguyen, D.N., Dutkiewicz, E. (2019). Optimal and Fast Real-Time Resource Slicing with Deep Dueling Neural Networks. *IEEE Journal on Selected Areas in Communications*, 37, 1455–1470.(https://doi.org/10.1109/JSAC.2019.2925399)
30. Van Huynh, N., Hoang, D.T., Nguyen, D.N., Dutkiewicz, E. (2019). Real-Time Network Slicing with Uncertain Demand: A Deep Learning Approach. In *Proceedings of the ICC 2019-2019 IEEE International Conference on Communications*, Shanghai, China, 20–24 May 2019, pp. 1–6.
31. Sun, G., Xiong, K., Boateng, G.O., Ayepah-Mensah, D., Liu, G., Jiang, W. (2019). Autonomous Resource Provisioning and Resource Customization for Mixed Traffics in Virtualized Radio Access Network. *IEEE Systems Journal*, 13, 2454–2465. (https://doi.org/10.1109/JSYST.2018.2875543)

32. Troia, S., Alvizu, R., Maier, G. (2019). Reinforcement Learning for Service Function Chain Reconfiguration in NFV-SDN Metro-Core Optical Networks. *IEEE Access*, 7, 167944–167957. (https://doi.org/10.1109/ACCESS.2019.2947800)
33. TIM. Big Data Challenge. (2014). Available online: (http://theodi.fbk.eu/openbigdata/)
34. Nassar, A., Yilmaz, Y. (2022). Deep Reinforcement Learning for Adaptive Network Slicing in 5G for Intelligent Vehicular Systems and Smart Cities. *IEEE Internet of Things Journal*, 9, 222–235. (https://doi.org/10.1109/JIOT.2021.3066102)
35. Hua, Y., Li, R., Zhao, Z., Chen, X., Zhang, H. (2020). GAN-Powered Deep Distributional Reinforcement Learning for Resource Management in Network Slicing. *IEEE Journal on Selected Areas in Communications*, 38, 334–349. (https://doi.org/10.1109/JSAC.2019.2947867)
36. Li, R., Wang, C., Zhao, Z., Guo, R., Zhang, H. (2020). The LSTM-Based Advantage Actor-Critic Learning for Resource Management in Network Slicing with User Mobility. *IEEE Communications Letters*, 24, 2005–2009. (https://doi.org/10.1109/LCOMM.2020.3000431)
37. Liu, Y., Ding, J., Zhang, Z.L., Liu, X. (2021). CLARA: A Constrained Reinforcement Learning Based Resource Allocation Framework for Network Slicing. In *Proceedings of the 2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 15–18 December 2021, pp. 1427–1437.
38. Abiko, Y., Saito, T., Ikeda, D., Ohta, K., Mizuno, T., Mineno, H. (2020a). Flexible Resource Block Allocation to Multiple Slices for Radio Access Network Slicing Using Deep Reinforcement Learning. *IEEE Access*, 8, 68183–68198. (https://doi.org/10.1109/ACCESS.2020.2989731)
39. Abiko, Y., Saito, T., Ikeda, D., Ohta, K., Mizuno, T., Mineno, H. (2020b). Radio Resource Allocation Method for Network Slicing using Deep Reinforcement Learning. In *Proceedings of the 2020 International Conference on Information Networking*, Barcelona, Spain, 7–10 January 2020, pp. 420–425.
40. Abiko, Y., Mochizuki, D., Saito, T., Ikeda, D., Mizuno, T., Mineno, H. (2019). Proposal of Allocating Radio Resources to Multiple Slices in 5G using Deep Reinforcement Learning. In *Proceedings of the 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, Osaka, Japan, 15–18 October 2019, pp. 1–2.
41. Yang, G., Liu, Q., Zhou, X., Qian, Y., Wu, W. (2019). Two-Tier Resource Allocation in Dynamic Network Slicing Paradigm with Deep Reinforcement Learning. In *Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, 9–13 December 2019, pp. 1–6.
42. Xi, R., Chen, X., Chen, Y., Li, Z. (2019). Real-Time Resource Slicing for 5G RAN via Deep Reinforcement Learning. In *Proceedings of the 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, Tianjin, China, 4–6 December 2019, pp. 625–632.
43. Koo, J., Mendiratta, V.B., Rahman, M.R., Walid, A. (2019). Deep Reinforcement Learning for Network Slicing with Heterogeneous Resource Requirements and Time Varying Traffic Dynamics. In *Proceedings of the 2019 15th International Conference on Network and Service Management (CNSM)*, Halifax, NS, Canada, 1–25 October 2019, pp. 1–5.
44. Statistical Workload Injector for Mapreduce (Swim). Available online: [Link](https://github.com/SWIMProjectUCB/SWIM/wiki)
45. Meng, S., Wang, Z., Ding, H., Wu, S., Li, X., Zhao, P., Zhu, C., Wang, X. (2019). RAN Slice Strategy Based on Deep Reinforcement Learning for Smart Grid. In *Proceedings of the 2019 Computing, Communications and IoT Applications*, Shenzhen, China, 26–28 October 2019, pp. 6–11.

46. Sun, G., Gebrekidan, Z.T., Boateng, G.O., Ayepah-Mensah, D., Jiang, W. (2019). Dynamic Reservation and Deep Reinforcement Learning Based Autonomous Resource Slicing for Virtualized Radio Access Networks. *IEEE Access*, 7, 45758–45772. (https://doi.org/10.1109/ACCESS.2019.2923432)
47. Sun, G., Zemuy, G.T., Xiong, K. (2018). Dynamic Reservation and Deep Reinforcement Learning based Autonomous Resource Management for wireless Virtual Networks. In *Proceedings of the 2018 IEEE 37th International Performance Computing and Communications Conference*, Orlando, FL, USA, 17–19 November 2018, pp. 1–4.
48. Sun, G., Al-Ward, H., Boateng, G.O., Liu, G. (2019). Autonomous Cache Resource Slicing and Content Placement at Virtualized Mobile Edge Network. *IEEE Access*, 7, 84727–84743. (https://doi.org/10.1109/ACCESS.2019.2925080)
49. Xu, Y., Zhao, Z., Cheng, P., Chen, Z., Ding, M., Vucetic, B., Li, Y. (2021). Constrained Reinforcement Learning for Resource Allocation in Network Slicing. *IEEE Communications Letters*, 25, 1554–1558. (https://doi.org/10.1109/LCOMM.2021.3066200)
50. Filali, A., Mlika, Z., Cherkaoui, S., Kobbane, A. (2022). Dynamic SDN-based Radio Access Network Slicing with Deep Reinforcement Learning for URLLC and eMBB Services. *IEEE Transactions on Network Science and Engineering*, 1–14. (https://doi.org/10.1109/TNSE.2021.3143321)
51. Schneider, S., Satheeschandran, N. P., Peuster, M., & Karl, H. (2020, June). Machine learning for dynamic resource allocation in network function virtualization. In 2020 6th IEEE Conference on Network Softwarization (NetSoft) (pp. 122-130). IEEE.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.