Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(6), 135-149
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

AI-Driven CPU Resource Management in Cloud Operating Systems

Yihan Wang[1], Suchuan Xing[2*]

[1]School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
Email: *sx80@alumni.duke.edu

How to cite this paper: Wang, Y.H. and Xing, S.C. (2025) AI-Driven CPU Resource Management in Cloud Operating Systems. *Journal of Computer and Communications*, **13**, 135-149.
https://doi.org/10.4236/jcc.2025.136009

Received: May 19, 2025
Accepted: June 21, 2025
Published: June 24, 2025

Abstract

As cloud computing continues to grow, efficient resource management remains a critical aspect of maintaining high-performance and cost-effective cloud infrastructures. CPU resource management in cloud operating systems is a key component that influences the scalability, efficiency, and responsiveness of cloud services. Traditional methods of resource allocation often struggle to adapt to the dynamic nature of cloud environments, where workloads can fluctuate in real time. This paper explores the integration of Artificial Intelligence (AI) in optimizing CPU resource management in cloud systems. By leveraging AI techniques such as machine learning and reinforcement learning, cloud operating systems can automate resource allocation, predict demand, and optimize system performance. The paper discusses the role of AI in enhancing resource efficiency, scalability, and fault tolerance, while also addressing the challenges of implementing AI-driven solutions. The potential of AI to transform the way cloud operating systems manage CPU resources offers exciting possibilities for future developments in cloud technology.

Keywords

CPU Resource Management, Cloud Computing, Machine Learning, Cloud Operating Systems, Resource Allocation

1. Introduction

Cloud computing has revolutionized the way businesses operate by providing on-demand access to computing resources, allowing organizations to scale their infrastructure without significant upfront investment [1]. However, the complexity of managing cloud resources—especially CPU resources—has posed significant challenges [2]. Cloud environments are inherently dynamic, with fluctuating work-

loads and varying demands on processing power. Traditional CPU resource management techniques, often static and reliant on predefined configurations, are ill-suited for adapting to these variations in real time [3].

The growing adoption of virtualization and the increasing number of cloud applications have exacerbated the demand for more efficient and dynamic resource management [4]. Traditional methods, such as manual allocation and static load balancing, can lead to underutilization of resources or, conversely, over-provisioning, which leads to wasted computing power and higher operational costs [5]. Additionally, as cloud infrastructures become more complex, managing resource allocation manually becomes increasingly challenging, especially for large-scale systems.

Artificial Intelligence (AI) offers a powerful solution to these challenges by enabling automated decision-making for resource management [6]. Machine learning algorithms, especially those based on reinforcement learning, provide the ability to predict demand, adjust resource allocation dynamically, and optimize CPU usage in real time [7]. These AI techniques not only improve resource efficiency, but also contribute to scalability, fault tolerance, and energy efficiency—crucial aspects in today's cloud computing landscape [8].

This paper explores the role of AI-driven CPU resource management in cloud operating systems, discussing how AI techniques can optimize performance, enhance scalability, and overcome the limitations of traditional resource management methods. It will also examine real-world applications of AI in cloud infrastructures, the benefits and challenges of integrating AI into resource management, and the future directions for research and development in this area.

2. Traditional CPU Resource Management in Cloud Systems

In traditional cloud computing environments, CPU resource management has typically relied on static configurations, where resources are allocated based on predefined policies or system requirements [9]. Early cloud systems would allocate resources manually or use basic load-balancing techniques to distribute tasks across servers [10]. However, as cloud environments evolved to support larger and more complex workloads, these methods began to show limitations [11].

The primary challenge with traditional CPU resource management lies in its inability to adapt dynamically to the changing nature of workloads [12]. Virtualization technologies, such as hypervisors and Virtual Machines (VMs), allowed for the efficient sharing of hardware resources across multiple tenants, but these solutions often relied on over-provisioning or manual intervention to handle resource allocation [13]. For instance, a cloud administrator might manually set CPU limits or adjust virtual machine configurations, but this does not account for real-time changes in resource demand, leading to inefficiencies such as underutilized CPUs during low-demand periods or CPU contention during peak load times.

Additionally, traditional CPU management systems often struggle with load

balancing in large-scale, distributed cloud infrastructures [14]. While basic load balancers direct incoming traffic to servers, they may not always optimize CPU usage efficiently, especially in environments with high variability in workload patterns. The lack of predictive resource management can lead to issues such as overload situations where servers are unable to meet processing demands or idle resources that remain underutilized during periods of low demand.

To address these issues, cloud providers began implementing more automated resource allocation systems, which use monitoring tools to assess CPU usage and dynamically adjust resources based on performance metrics [15]. These systems rely on metrics such as CPU utilization, memory usage, and disk I/O to make real-time decisions about allocating resources. However, even these more advanced solutions often fall short when handling highly variable workloads or dealing with complex data patterns that require more sophisticated decision-making [16].

In summary, traditional CPU resource management methods in cloud systems, while effective for basic applications, lack the flexibility and efficiency needed for modern cloud environments [17]. With the increasing demand for elasticity, scalability, and real-time performance optimization, the limitations of these approaches have become more apparent, paving the way for the integration of AI-driven resource management systems.

3. AI for CPU Resource Optimization

The integration of AI into CPU resource management offers the potential to solve many of the challenges faced by traditional systems [18]. Unlike conventional methods, AI-driven approaches can learn from data, predict future resource requirements, and adjust resource allocation dynamically, ensuring optimal performance and efficiency [19]. By using techniques such as Machine Learning (ML), Reinforcement Learning (RL), and predictive analytics, AI models are capable of adapting to real-time changes in workload demands and optimizing CPU usage without human intervention [20].

RL is one of the most promising AI techniques for optimizing CPU resource management [21]. In an RL-based system, an agent learns to make decisions by interacting with the environment and receiving feedback in the form of rewards or penalties. The agent's objective is to maximize cumulative rewards by taking actions that optimize system performance, such as allocating CPU resources effectively. In the context of cloud systems, RL can be used to dynamically allocate CPU resources to different Virtual Machines (VMs) or containers based on workload demands. For example, when a server experiences high CPU utilization, the RL agent can move resources to other underutilized servers or provision additional VMs, ensuring that the system remains balanced and efficient.

ML techniques are also widely used in AI-driven CPU optimization [22]. Supervised learning algorithms, such as decision trees and random forests, can be used to analyze historical data and predict future resource demands. By training on past workload patterns, ML models can predict periods of high or low CPU usage and

preemptively allocate resources accordingly. For example, during periods of high traffic, the system can anticipate the need for more CPU power and allocate additional virtual machines or increase CPU cores to meet demand. This predictive capability ensures that resources are used efficiently without the need for constant manual intervention.

Another important AI-driven approach to CPU optimization is predictive analytics [23]. By analyzing large datasets in real time, predictive models can identify patterns in workload behavior and adjust resources dynamically. For instance, predictive models can forecast traffic spikes, server downtimes, or increased computational needs based on historical usage trends and external factors such as time of day or market conditions [24]. This allows cloud providers to proactively manage resources, ensuring that CPU power is allocated in anticipation of future needs rather than as a reactive response to problems.

One of the key advantages of AI-driven CPU resource management is its ability to operate autonomously without human intervention [25]. In large-scale cloud environments, where workloads can vary dramatically throughout the day or even minute by minute, relying on automated AI systems to manage CPU resources enables more efficient operations and scalability. For example, during periods of high demand, AI systems can automatically allocate more CPU cores or scale out services by spinning up additional instances [26]. During low-demand periods, the system can automatically scale down resources, reducing waste and improving cost efficiency.

In addition to improving efficiency, AI-driven CPU resource management also contributes to fault tolerance and high availability in cloud systems [27]. By continuously monitoring the system and adjusting resource allocation in real time, AI can help mitigate the impact of system failures or hardware malfunctions. If a server goes down or experiences issues, AI systems can detect the problem immediately and redistribute CPU resources across other servers, minimizing service interruptions and ensuring that cloud applications remain available to users.

AI-driven approaches to CPU resource optimization also help in managing multi-tenant environments and virtualized infrastructures, where multiple users or applications share the same physical resources [28]. Machine learning models can prioritize resource allocation based on the importance or priority of specific workloads, ensuring that critical applications receive the resources they need without compromising performance for other tenants.

Table 1 provides a detailed comparative analysis of the primary AI techniques employed in CPU resource management, revealing significant variations in computational complexity, training requirements, and performance characteristics. The comparison highlights that while Deep Q-Networks and Transformer models offer high accuracy with $O(n^2)$ complexity, they require substantially longer training times and higher inference latency compared to Random Forest algorithms, which provide moderate accuracy with superior interpretability and faster response times.

Table 1. Comparison of AI techniques for CPU resource management.

AI Technique	Primary Use Case	Training Time	Accuracy	Interpretability
Deep Q-Network (DQN)	Dynamic Resource Allocation	24 - 48 hours	High	Low
LSTM Networks	Demand Prediction	12 - 24 hours	High	Medium
Transformer Models	Pattern Recognition	36 - 72 hours	Very High	Medium
Random Forest	Classification	2 - 6 hours	Medium	High
Genetic Algorithm	Optimization	6 - 12 hours	Medium	High

4. Applications and Benefits of AI-Driven CPU Resource Management

AI-driven CPU resource management offers numerous advantages in cloud operating systems, improving efficiency, scalability, and fault tolerance [29]. By leveraging ML and RL techniques, these systems can automate resource allocation, predict CPU demands, and dynamically adjust resources in real time. The integration of AI into cloud resource management not only enhances system performance but also provides significant cost savings and ensures that resources are used optimally across different workloads.

One of the most significant benefits of AI-driven CPU resource management is its ability to automate and optimize resource allocation without human intervention [30]. Traditional resource management methods often rely on static configurations or manual adjustments by cloud administrators. These methods are time-consuming and may lead to either over-provisioning or under-provisioning of resources, resulting in inefficiencies. AI systems, on the other hand, can continuously monitor workloads and automatically adjust resource allocation based on real-time demands, ensuring that CPU resources are always available when needed and not wasted during periods of low demand.

For example, in cloud environments supporting large-scale applications, workloads can vary significantly depending on factors such as user demand, time of day, or seasonal patterns [31]. RL can optimize resource allocation by learning from past experiences and making decisions to maximize performance. The AI agent, for instance, could learn the most efficient way to allocate CPU resources based on workload patterns, minimizing delays or bottlenecks. During periods of peak demand, RL agents can spin up additional VMs or scale CPU cores to meet performance requirements, ensuring that user experience remains smooth [32]. Similarly, during off-peak times, AI can reduce the number of active VMs, thereby reducing resource wastage and optimizing costs.

Another major benefit is the predictive capabilities of AI systems. Machine learning algorithms can be trained on historical data to forecast future CPU demands, identifying potential traffic spikes or periods of high computation needs [33]. These predictive capabilities allow cloud systems to proactively allocate resources before

demand increases, ensuring that sufficient CPU power is available without waiting for resource shortages to impact performance. For example, predictive analytics can forecast a sudden spike in traffic to a website or application and preemptively allocate additional CPU resources, ensuring that performance does not degrade when demand peaks.

Furthermore, AI-driven CPU resource management enhances fault tolerance and system reliability [34]. Cloud environments often experience failures due to hardware malfunctions, network issues, or unexpected traffic surges. With AI systems in place, resource allocation can be dynamically adjusted in response to failures. If a server or VM experiences issues, the AI system can detect the problem immediately and reallocate resources across other servers or VMs. This automated recovery ensures that services remain available to users even during system failures, contributing to high availability in cloud infrastructures.

In multi-tenant cloud environments, where multiple users share the same physical resources, AI systems can manage resource allocation based on workload prioritization [35]. Machine learning models can identify which workloads are mission-critical and ensure that these workloads receive the CPU resources they need, while less critical applications are allocated fewer resources. This resource prioritization ensures that high-priority tasks, such as financial transactions or medical data processing, maintain high performance without being interrupted by lower-priority applications.

To guide practical implementation decisions, **Table 2** provides recommendations for AI technique selection based on cloud environment characteristics. The table shows that smaller enterprises can benefit from simpler approaches like Random Forest algorithms, while hyperscale deployments require more sophisticated Multi-Agent Reinforcement Learning systems to manage complex distributed resources effectively.

Overall, AI-driven CPU resource management not only improves the efficiency and scalability of cloud systems but also provides the flexibility needed to meet the demands of dynamic and growing cloud environments [29]. The ability to automate resource allocation, predict demand, and ensure fault tolerance makes AI-driven systems ideal for modern cloud infrastructures, where performance and cost optimization are essential.

Table 2. AI technique selection guide for different cloud environments.

Cloud Environment Type	Recommended AI Technique	Key Advantage	Implementation Difficulty
Small Enterprise (100 - 500 VMs)	Random Forest	Low complexity, fast deployment	Easy
Medium Enterprise (500 - 2000 VMs)	LSTM Networks	Good prediction accuracy	Moderate
Large Enterprise (2000 - 5000 VMs)	Deep Q-Network	Dynamic resource allocation	Moderate
Hyperscale (5000+ VMs)	Multi-Agent RL	Distributed decision making	High
High-Performance Computing	Transformer Models	Superior pattern recognition	High
Edge Computing	Lightweight ML Models	Low latency, minimal overhead	Easy

5. Ethical Considerations and Privacy Protection Mechanisms

Data privacy and security concerns represent critical challenges that require comprehensive mitigation strategies in AI-driven cloud resource management systems. This section details specific approaches to address ethical considerations and implement effective privacy protection mechanisms.

The comprehensive privacy protection framework required for ethical AI deployment is illustrated in Figure 1, which organizes privacy-preserving mechanisms into six interconnected components covering data protection, computation security, communication safety, access control, compliance, and transparency. This framework demonstrates how multiple privacy techniques must work in concert to address the complex privacy requirements of AI-driven cloud resource management systems while maintaining operational effectiveness.

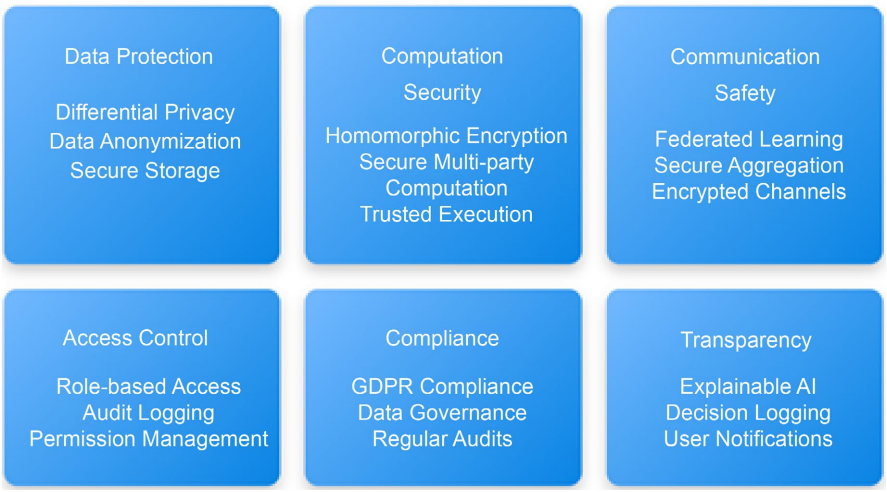

Figure 1. Privacy-preserving AI framework components.

5.1. Privacy-Preserving AI Techniques

Differential privacy mechanisms can be integrated into AI training processes to protect sensitive workload and user data [36]. The technique adds calibrated noise to training data or model outputs, ensuring that individual data points cannot be identified while maintaining model accuracy. The privacy budget parameter controls the trade-off between privacy and utility, allowing organizations to balance protection requirements with system performance needs.

Federated learning architectures enable cloud resource management systems to train models across distributed data centers without centralizing sensitive data [37]. Each data center trains local models on private data, sharing only encrypted model updates with a central coordinator. The global model aggregation process ensures that individual data center information remains private while benefiting from collective learning across the entire cloud infrastructure.

Homomorphic encryption techniques enable computation on encrypted data

without requiring decryption [38]. This allows AI models to process sensitive resource utilization data while maintaining privacy throughout the computation process. Partially homomorphic encryption schemes support essential mathematical operations on encrypted values, enabling secure aggregation of performance metrics across multiple tenants without exposing individual usage patterns.

5.2. Fairness and Bias Mitigation Strategies

Algorithmic fairness metrics ensure equitable resource allocation across different tenant categories [39]. Demographic parity requires that resource allocation probabilities remain consistent across tenant groups, while equalized opportunity ensures that tenants with similar priority levels receive similar treatment regardless of their category. Individual fairness requires that similar tenants receive similar resource allocations based on defined similarity metrics that reflect legitimate business requirements.

Bias detection and correction processes involve regular auditing to identify potential discriminatory patterns in resource allocation decisions [40]. Statistical tests compare allocation patterns across different tenant categories, while fairness-aware machine learning algorithms incorporate bias correction during the training process. Reweighting techniques adjust training sample weights to balance representation across different groups, adversarial debiasing methods train specialized networks to minimize discriminatory patterns, and post-processing calibration adjusts model outputs to satisfy fairness constraints.

5.3. Transparency and Explainability Mechanisms

Explainable AI techniques address the black-box nature of complex AI models by providing interpretable explanations for resource allocation decisions [41]. Local Interpretable Model-agnostic Explanations generate local explanations by training interpretable models around specific prediction instances. For resource allocation decisions, this approach can identify which workload characteristics most influenced the allocation decision, providing transparency to system administrators and tenants.

Shapley Additive Explanations provide consistent and accurate feature attribution using cooperative game theory principles [42]. This method explains how each input feature contributes to the final resource allocation decision, enabling stakeholders to understand the reasoning behind automated allocation choices. The approach ensures that explanation values sum to the difference between the actual prediction and the expected baseline prediction.

For Transformer-based models, attention weight visualization shows which historical patterns most influence current predictions, providing insights into model decision-making processes. This visualization capability helps system administrators understand how the AI system weighs different factors when making resource allocation decisions, building trust and enabling better system monitoring.

5.4. Governance and Compliance Framework

Comprehensive data governance frameworks establish clear policies for data collection, processing, and retention in AI-driven systems. Data minimization principles ensure that systems collect only necessary data for resource management purposes [43], while purpose limitation restricts data usage to specified resource optimization objectives. Storage limitation policies implement automatic data deletion after defined retention periods, and role-based access controls restrict data access with comprehensive audit logging.

Regulatory compliance integration ensures that AI systems meet requirements under relevant regulations such as GDPR, CCPA, and sector-specific standards [44]. Right-to-explanation provisions provide clear explanations for automated resource allocation decisions, data portability features enable tenants to export their resource usage data, consent management systems obtain explicit consent for data processing activities, and regular compliance audits systematically review data processing activities and AI decision-making processes.

Ethical AI review boards consisting of interdisciplinary committees review AI system designs [45], evaluate potential ethical implications, and provide ongoing oversight of AI-driven resource management systems. These boards include technical experts, ethicists, legal professionals, and stakeholder representatives who ensure that AI implementations align with organizational values and societal expectations.

Incident response and remediation procedures address potential ethical violations or privacy breaches through comprehensive response protocols. These procedures include immediate system isolation capabilities, forensic analysis tools and processes, stakeholder notification systems, and structured remediation planning frameworks. Regular drills and simulations ensure response readiness and enable continuous improvement of protection mechanisms based on lessons learned from exercises and real incidents.

6. Challenges and Limitations of AI-Driven CPU Resource Management

While the benefits of AI-driven CPU resource management are clear, there are several challenges and limitations that must be addressed to ensure successful implementation and widespread adoption. These challenges stem from the complexity of integrating AI models into existing cloud infrastructures, the need for large and high-quality datasets, and the trade-offs between accuracy and real-time performance [46].

One of the primary challenges is data quality and availability. AI models require large, high-quality datasets to train effectively, but in cloud environments, data can be noisy, incomplete, or inconsistent [47]. The lack of clean, labeled data can lead to inaccurate predictions and suboptimal resource allocation. For instance, if a machine learning model is trained on incomplete data about system performance, it may fail to predict CPU demand accurately, leading to either over-provisioning

or under-provisioning of resources. Furthermore, cloud systems are often distributed across multiple locations, making it difficult to gather and standardize data in real time. Ensuring data consistency and quality is therefore crucial for the success of AI-driven resource management systems.

Another challenge is the computational overhead associated with implementing AI algorithms in real-time cloud environments. While AI models can greatly improve resource allocation, they also require significant computational resources for both training and execution [48]. This computational burden can introduce delays in decision-making, especially in environments with highly dynamic workloads. If the AI system requires a long time to process data and make decisions, it could result in performance bottlenecks that hinder overall system efficiency. This issue is particularly important in environments where low latency and real-time decision-making are crucial, such as in online financial services or e-commerce platforms.

In addition to computational overhead, model complexity remains a significant barrier. More complex machine learning models, such as deep learning networks, often provide high accuracy but are difficult to explain and interpret [49]. In contrast, simpler models may be more transparent but lack the predictive power necessary for optimizing CPU usage in large-scale cloud environments. The trade-off between accuracy and interpretability becomes a key concern in risk management, as complex models may yield better performance but lack the transparency required for effective decision-making. Balancing these two factors is a major challenge for AI-driven resource management systems.

Moreover, the integration of AI with legacy cloud systems presents another obstacle. Many cloud providers still rely on older systems and infrastructure that were not designed to accommodate machine learning-based approaches. Transitioning to AI-driven resource management requires significant changes in both hardware and software, as well as investment in training personnel to work with these new technologies. The integration complexity often leads to slow adoption rates, particularly for smaller cloud providers with limited resources.

Data privacy and security are also major concerns. As cloud systems increasingly rely on AI to manage sensitive data, it is essential to ensure that these systems comply with privacy regulations such as the GDPR. AI models must be designed with strong data security protocols to prevent unauthorized access or misuse of personal and financial data.

Finally, ethical concerns arise when using AI to make autonomous decisions about resource allocation, especially when it comes to prioritizing certain workloads over others. If AI models are trained on biased data, they may inadvertently discriminate against certain users or applications, leading to unfair resource distribution. Ensuring that AI systems are fair, transparent, and non-discriminatory is critical for maintaining trust in cloud environments.

7. Conclusions

In conclusion, AI-driven CPU resource management has the potential to transform

the way cloud operating systems manage resources, offering significant improvements in efficiency, scalability, and fault tolerance. By leveraging machine learning and reinforcement learning, cloud providers can automate the allocation of CPU resources, predict demand, and optimize system performance in real time. However, the implementation of AI-driven systems comes with challenges, including the need for high-quality data, computational overhead, model complexity, and integration with legacy systems.

Despite these challenges, the future of AI in CPU resource management looks promising. As AI techniques continue to evolve, the ability to balance accuracy and interpretability, while addressing ethical and privacy concerns, will be crucial in making AI systems more transparent and trustworthy. The integration of AI into cloud operating systems will play a key role in improving cloud infrastructure, ensuring that resources are allocated dynamically and efficiently to meet the needs of modern applications. As AI continues to develop, we can expect even more advanced resource management systems that are capable of adapting to the complexities of the cloud, offering new levels of performance and cost optimization for cloud providers and users alike.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sunyaev, A. and Sunyaev, A. (2020) Cloud Computing. In: Sunyaev, A., Ed., *Internet Computing*: *Principles of Distributed Systems and Emerging Internet-Based Technologies*, Springer, 195-236.

[2] Gill, S.S., Garraghan, P., Stankovski, V., Casale, G., Thulasiram, R.K., Ghosh, S.K., *et al.* (2019) Holistic Resource Management for Sustainable and Reliable Cloud Computing: An Innovative Solution to Global Challenge. *Journal of Systems and Software*, **155**, 104-129. https://doi.org/10.1016/j.jss.2019.05.025

[3] Thinakaran, P. (2021) Embracing Heterogeneity in Cloud Platforms through Adaptable Resource Management Framework. Ph.D. Thesis, The Pennsylvania State University.

[4] Olsen, D. and Anagnostopoulos, I. (2017) Performance-Aware Resource Management of Multi-Threaded Applications on Many-Core Systems. *Proceedings of the Great Lakes Symposium on VLSI* 2017, Banff, 10-12 May 2017, 119-124. https://doi.org/10.1145/3060403.3060426

[5] Oyediran, M.O., Ojo, O.S., Ajagbe, S.A., Aiyeniko, O., Chima Obuzor, P. and Adigun, M.O. (2024) Comprehensive Review of Load Balancing in Cloud Computing System. *International Journal of Electrical and Computer Engineering*, **14**, 3244-3255. https://doi.org/10.11591/ijece.v14i3.pp3244-3255

[6] Duan, Y., Edwards, J.S. and Dwivedi, Y.K. (2019) Artificial Intelligence for Decision Making in the Era of Big Data—Evolution, Challenges and Research Agenda. *International Journal of Information Management*, **48**, 63-71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

[7] Jiang, Y., Kodialam, M., Lakshman, T.V., Mukherjee, S. and Tassiulas, L. (2021) Resource Allocation in Data Centers Using Fast Reinforcement Learning Algorithms. *IEEE*

Transactions on Network and Service Management, **18**, 4576-4588. https://doi.org/10.1109/tnsm.2021.3100460

[8] Belgaum, M.R., Alansari, Z., Musa, S., Mansoor Alam, M. and Mazliham, M.S. (2021) Role of Artificial Intelligence in Cloud Computing, IoT and SDN: Reliability and Scalability Issues. *International Journal of Electrical and Computer Engineering*, **11**, 4458-4470. https://doi.org/10.11591/ijece.v11i5.pp4458-4470

[9] Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M. and Buyya, R. (2022) Machine Learning (ML)-Centric Resource Management in Cloud Computing: A Review and Future Directions. *Journal of Network and Computer Applications*, **204**, Article ID: 103405. https://doi.org/10.1016/j.jnca.2022.103405

[10] Kumar, P. and Kumar, R. (2019) Issues and Challenges of Load Balancing Techniques in Cloud Computing. *ACM Computing Surveys*, **51**, 1-35. https://doi.org/10.1145/3281010

[11] Menaka, M. and Sendhil Kumar, K.S. (2022) Workflow Scheduling in Cloud Environment—Challenges, Tools, Limitations & Methodologies: A Review. *Measurement: Sensors*, **24**, Article ID: 100436. https://doi.org/10.1016/j.measen.2022.100436

[12] Jennings, B. and Stadler, R. (2014) Resource Management in Clouds: Survey and Research Challenges. *Journal of Network and Systems Management*, **23**, 567-619. https://doi.org/10.1007/s10922-014-9307-7

[13] Thyagaturu, A.S., Shantharama, P., Nasrallah, A. and Reisslein, M. (2022) Operating Systems and Hypervisors for Network Functions: A Survey of Enabling Technologies and Research Studies. *IEEE Access*, **10**, 79825-79873. https://doi.org/10.1109/access.2022.3194913

[14] Suleiman, N. and Murtaza, Y. (2024) Scaling Microservices for Enterprise Applications: Comprehensive Strategies for Achieving High Availability, Performance Optimization, Resilience, and Seamless Integration in Large-Scale Distributed Systems and Complex Cloud Environments. *Applied Research in Artificial Intelligence and Cloud Computing*, **7**, 46-82.

[15] Moghaddam, S.K., Buyya, R. and Ramamohanarao, K. (2019) Performance-Aware Management of Cloud Resources: A Taxonomy and Future Directions. *ACM Computing Surveys*, **52**, 1-37. https://doi.org/10.1145/3337956

[16] Sharma, K., Salagrama, S., Parashar, D. and Chugh, R.S. (2024) AI-Driven Decision Making in the Age of Data Abundance: Navigating Scalability Challenges in Big Data Processing. *Revue d'Intelligence Artificielle*, **38**, 1335-1340. https://doi.org/10.18280/ria.380427

[17] Buyya, R., Ilager, S. and Arroba, P. (2023) Energy-Efficiency and Sustainability in New Generation Cloud Computing: A Vision and Directions for Integrated Management of Data Centre Resources and Workloads. *Software: Practice and Experience*, **54**, 24-38. https://doi.org/10.1002/spe.3248

[18] Gill, S.S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., *et al.* (2022) AI for Next Generation Computing: Emerging Trends and Future Directions. *Internet of Things*, **19**, Article ID: 100514. https://doi.org/10.1016/j.iot.2022.100514

[19] Ramamoorthi, V. (2024) AI-Driven Cloud Resource Optimization Framework for Real-Time Allocation. *Journal of Advanced Computing Systems*, **1**, 8-15. https://doi.org/10.69987/jacs.2021.10102

[20] Kalusivalingam, A.K., Sharma, A., Patel, N. and Singh, V. (2022) Leveraging Reinforcement Learning and Genetic Algorithms for Enhanced Cloud Infrastructure Optimization. *International Journal of AI and ML*, **3**, 1-25.

[21] Joloudari, J.H., Alizadehsani, R., Nodehi, I., Mojrian, S., Fazl, F., Shirkharkolaie, S.K. and Acharya, U.R. (2022) Resource Allocation Optimization Using Artificial Intelligence Methods in Various Computing Paradigms: A Review. arXiv: 2203.12315.

[22] Biswas, P., Rashid, A., Biswas, A., Nasim, M.A.A., Chakraborty, S., Gupta, K.D., *et al.* (2024) AI-Driven Approaches for Optimizing Power Consumption: A Comprehensive Survey. *Discover Artificial Intelligence*, **4**, Article No. 116. https://doi.org/10.1007/s44163-024-00211-7

[23] Nama, P., Pattanayak, S. and Meka, H.S. (2023) AI-Driven Innovations in Cloud Computing: Transforming Scalability, Resource Management, and Predictive Analytics in Distributed Systems. *International Research Journal of Modernization in Engineering Technology and Science*, **5**, 4165-4174.

[24] Toumi, H., Brahmi, Z. and Gammoudi, M.M. (2022) RTSLPS: Real Time Server Load Prediction System for the Ever-Changing Cloud Computing Environment. *Journal of King Saud University—Computer and Information Sciences*, **34**, 342-353. https://doi.org/10.1016/j.jksuci.2019.12.004

[25] Banerjee, S. (2025) Intelligent Cloud Systems: AI-Driven Enhancements in Scalability and Predictive Resource Management. (Preprints) https://doi.org/10.20944/preprints202501.1153.v1

[26] Ilager, S., Muralidhar, R. and Buyya, R. (2020) Artificial Intelligence (AI)-Centric Management of Resources in Modern Distributed Computing Systems. 2020 *IEEE Cloud Summit*, Harrisburg, 21-22 October 2020, 1-10. https://doi.org/10.1109/ieeecloudsummit48914.2020.00007

[27] Bhattarai, A. (2023) AI-Enhanced Cloud Computing: Comprehensive Review of Resource Management, Fault Tolerance, and Security. *Emerging Trends in Machine Intelligence and Big Data*, **15**, 39-50.

[28] Govindarajan, V., Sonani, R. and Patel, P.S. (2020) Secure Performance Optimization in Multi-Tenant Cloud Environments. *Annals of Applied Sciences*, **1**, 1-9.

[29] Banerjee, S. (2024) Intelligent Cloud Systems: AI-Driven Enhancements in Scalability and Predictive Resource Management. *International Journal of Advanced Research in Science, Communication and Technology*, **4**, 266-276.

[30] Zheng, H., Xu, K., Zhang, M., Tan, H. and Li, H. (2024) Efficient Resource Allocation in Cloud Computing Environments Using AI-Driven Predictive Analytics. *Applied and Computational Engineering*, **82**, 17-23. https://doi.org/10.54254/2755-2721/82/2024glg0055

[31] Singh, P., Gupta, P., Jyoti, K. and Nayyar, A. (2019) Research on Auto-Scaling of Web Applications in Cloud: Survey, Trends and Future Directions. *Scalable Computing: Practice and Experience*, **20**, 399-432. https://doi.org/10.12694/scpe.v20i2.1537

[32] Read, M.R., Dehury, C., Srirama, S.N. and Buyya, R. (2024) Deep Reinforcement Learning (DRL)-Based Methods for Serverless Stream Processing Engines: A Vision, Architectural Elements, and Future Directions. In: Mukherjee, A., De, D. and Buyya, R., Eds., *Resource Management in Distributed Systems*, Springer Nature, 285-314. https://doi.org/10.1007/978-981-97-2644-8_14

[33] Aqib, M., Mehmood, R., Alzahrani, A., Katib, I., Albeshri, A. and Altowaijri, S.M. (2019) Smarter Traffic Prediction Using Big Data, In-Memory Computing, Deep Learning and GPUs. *Sensors*, **19**, Article No. 2206. https://doi.org/10.3390/s19092206

[34] Kumar, P. and Nair, K. (2024) Fault Tolerance and Reliability in Advanced Compu-ting Systems: Techniques and Trends. *Journal of Advanced Computing Systems*, **4**, 19-25.

[35] Ameur, A.B. (2023) Artificial Intelligence for Resource Allocation in Multi-Tenant Edge Computing. Doctoral Dissertation, Institut Polytechnique de Paris.

[36] Zhu, T., Ye, D., Wang, W., Zhou, W. and Yu, P.S. (2022) More than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence. *IEEE Transactions on Knowledge and Data Engineering*, **34**, 2824-2843. https://doi.org/10.1109/tkde.2020.3014246

[37] Drainakis, G., Pantazopoulos, P., Katsaros, K.V., Sourlas, V., Amditis, A. and Kaklamani, D.I. (2023) From Centralized to Federated Learning: Exploring Performance and End-to-End Resource Consumption. *Computer Networks*, **225**, Article ID: 109657. https://doi.org/10.1016/j.comnet.2023.109657

[38] Acar, A., Aksu, H., Uluagac, A.S. and Conti, M. (2018) A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Computing Surveys*, **51**, 1-35. https://doi.org/10.1145/3214303

[39] Jo, N., Tang, B., Dullerud, K., Aghaei, S., Rice, E. and Vayanos, P. (2023) Fairness in Contextual Resource Allocation Systems: Metrics and Incompatibility Results. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 11837-11846. https://doi.org/10.1609/aaai.v37i10.26397

[40] Murikah, W., Nthenge, J.K. and Musyoka, F.M. (2024) Bias and Ethics of AI Systems Applied in Auditing—A Systematic Review. *Scientific African*, **25**, e02281. https://doi.org/10.1016/j.sciaf.2024.e02281

[41] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., *et al.* (2023) Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, **16**, 45-74. https://doi.org/10.1007/s12559-023-10179-8

[42] Li, M., Sun, H., Huang, Y. and Chen, H. (2024) Shapley Value: From Cooperative Game to Explainable Artificial Intelligence. *Autonomous Intelligent Systems*, **4**, Article No. 2. https://doi.org/10.1007/s43684-023-00060-8

[43] Abu-Elkheir, M., Hayajneh, M. and Ali, N. (2013) Data Management for the Internet of Things: Design Primitives and Solution. *Sensors*, **13**, 15582-15612. https://doi.org/10.3390/s131115582

[44] Ijaiya, H. and Odumuwagun, O.O. (2024) Advancing Artificial Intelligence and Safeguarding Data Privacy: A Comparative Study of EU and US Regulatory Frameworks Amid Emerging Cyber Threats. *International Journal of Research Publication and Reviews*, **5**, 3357-3375.

[45] Jordan, S.R. (2019) Designing Artificial Intelligence Review Boards: Creating Risk Metrics for Review of AI. 2019 *IEEE International Symposium on Technology and Society* (*ISTAS*), Medford, 15-16 November 2019, 1-7. https://doi.org/10.1109/istas48451.2019.8937942

[46] Aminizadeh, S., Heidari, A., Dehghan, M., Toumaj, S., Rezaei, M., Jafari Navimipour, N., *et al.* (2024) Opportunities and Challenges of Artificial Intelligence and Distributed Systems to Improve the Quality of Healthcare Service. *Artificial Intelligence in Medicine*, **149**, Article ID: 102779. https://doi.org/10.1016/j.artmed.2024.102779

[47] Liang, W., Tadesse, G.A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., *et al.* (2022) Advances, Challenges and Opportunities in Creating Data for Trustworthy AI. *Nature Machine Intelligence*, **4**, 669-677. https://doi.org/10.1038/s42256-022-00516-1

[48] Chen, C., Zhang, P., Zhang, H., Dai, J., Yi, Y., Zhang, H., *et al.* (2020) Deep Learning on Computational-Resource-Limited Platforms: A Survey. *Mobile Information Systems*, **2020**, Article ID: 8454327. https://doi.org/10.1155/2020/8454327

[49] Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., *et al.* (2022) Interpretable Deep

Learning: Interpretation, Interpretability, Trustworthiness, and Beyond. *Knowledge and Information Systems*, **64**, 3197-3234.
https://doi.org/10.1007/s10115-022-01756-8