JOURNAL OF ASIAN SCIENTIFIC RESEARCH (JOASR)
2025, Vol. 15, No. 3, May-June, pp. 13-18.
website: https://joasr.com

Multilayered AI Workload Optimizers for Cross-Platform Data Centers Supporting Dynamic Resource Allocation

Berdymyrat Gromov,
Research Scientist, Latvia.

Abstract

Modern data centers are undergoing a transformative shift as they adapt to dynamic workloads spanning heterogeneous platforms such as edge, fog, and cloud systems. Artificial Intelligence (AI)-driven workload optimizers are emerging as critical tools for dynamic resource allocation, enabling energy efficiency, cost reduction, and improved quality of service (QoS). This paper introduces a multilayered AI workload optimization framework tailored for cross-platform data centers. The framework incorporates real-time decision-making, predictive modeling, and reinforcement learning mechanisms to allocate resources adaptively across distributed infrastructures. By drawing upon advancements, we assess the challenges and performance implications of dynamic workload optimization, culminating in a hybrid AI system that balances scalability and responsiveness.

Keywords: AI workload optimization, cross-platform data centers, dynamic resource allocation, multilayer architecture, reinforcement learning, edge computing, cloud orchestration

Citation: Gromov B. (2025). Multilayered AI Workload Optimizers for Cross-Platform Data Centers Supporting Dynamic Resource Allocation. *Journal of Asian Scientific Research (JOASR),* 15(3), 13–18.

1. Introduction

Modern digital ecosystems rely heavily on data centers to execute computational tasks at scale. With the advent of the Internet of Things (IoT), machine learning, and high-frequency user interactions, the demand for responsive and adaptive data center architectures has surged. Traditional static provisioning methods are no longer sufficient due to increasing variability in workload intensity and diversity in hardware platforms.

In this context, AI-driven workload optimization is not only relevant but necessary. When integrated within multilayered architectures—encompassing edge, fog, and cloud systems—AI models can enable real-time, cross-platform workload distribution. This paper investigates such architectures and presents a comprehensive AI-based framework for dynamic resource allocation.

2. Literature Review

2.1 Early Developments in Workload Optimization

Workload optimization strategies initially focused on static scheduling and heuristic-based load balancing. Approaches like First-Come-First-Serve (FCFS) and Round Robin were dominant until early 2010s. However, as data traffic grew, more complex algorithms emerged. For example, Beloglazov et al. (2012) introduced energy-aware scheduling in cloud environments, using heuristics to reduce power consumption.

CloudSim, an early simulation toolkit, was widely used to test optimization algorithms, revealing significant trade-offs between energy efficiency and response time. In parallel, virtualization technologies improved resource abstraction, allowing workloads to be decoupled from specific hardware.

2.2 Rise of AI and Reinforcement Learning in Data Center Management

By the mid-2010s, AI began making inroads into data center optimization. Google's DeepMind application in 2016 to reduce cooling energy by 40% demonstrated the potential of reinforcement learning (RL) for real-time control systems. Later, model-based RL and deep Q-networks (DQNs) were adapted for dynamic resource allocation, offering better adaptability to workload fluctuations.

More recently, federated learning and distributed AI models have been explored for managing edge-to-cloud orchestration, especially in smart city and IoT contexts. These studies laid the groundwork for building multi-layered, cross-platform optimizers using AI techniques.

3. Architecture of Multilayered AI Workload Optimizer

3.1 Conceptual Framework

The proposed architecture integrates three tiers: edge, fog, and cloud, with each tier hosting a layer-specific optimizer that contributes to a global decision-making model. At the edge, lightweight AI agents perform rapid inference using on-device metrics. Fog layers perform data aggregation and intermediate optimization, while cloud layers manage global resource policies and predictive analytics.

Each layer communicates via a federated reinforcement learning model, enabling local decisions to be informed by global objectives. This distributed learning approach mitigates latency and improves robustness against node failures or data delays.

Figure 1. Multilayered AI Workload Optimizer Architecture

3.2 Decision-making Workflow

In the operational pipeline, workloads are tagged with context-specific metadata (e.g., latency sensitivity, memory demands). These tags are fed into the optimizer agents, which predict the optimal placement tier using a trained policy gradient model. The edge agents prioritize real-time response, fog nodes optimize for throughput, and cloud agents maximize energy efficiency.

This modular decision-making structure allows partial autonomy while maintaining alignment through a shared objective function. Model training is performed offline using historical workload traces and is fine-tuned periodically using online learning.

4. Dynamic Resource Allocation Algorithms

4.1 Reinforcement Learning Model

The core of the workload optimization system is a deep reinforcement learning (DRL) framework. The system state is represented by resource availability, task queue metrics, and network bandwidth. Actions involve assigning workloads to specific tiers. Rewards are computed based on task completion time, resource utilization efficiency, and energy overhead.

We implemented a Double DQN (DDQN) approach to address overestimation issues commonly encountered in standard DQNs. The model was trained on Google Cluster workload traces and validated using simulation environments based on CloudSim Plus.

4.2 Load Prediction and Proactive Scaling

To improve allocation accuracy, we integrated a time-series forecasting model (ARIMA and LSTM) to predict future workload demands. This enables proactive scaling and pre-emptive resource provisioning. The model forecasts 15-minute workload intervals with 87.3% accuracy (measured using MAPE – Mean Absolute Percentage Error).

Table 1. Workload Forecast Accuracy Comparison

Model	**MAPE (%)**	**RMSE**
ARIMA	14.2	5.6
LSTM	**12.7**	**4.9**
Moving Avg	22.3	8.1

5. Evaluation and Performance Metrics

5.1 Experimental Setup

We tested the proposed system in a hybrid testbed with Raspberry Pi edge devices, Jetson Nano fog nodes, and a virtualized OpenStack-based cloud environment. Workloads were emulated using Apache JMeter and customized scripts simulating sensor data, video analytics, and user queries.

Performance was evaluated based on average task latency, energy consumption, and resource utilization rate. Compared with baseline static allocation, our optimizer improved average task latency by 34% and reduced energy usage by 28%.

5.2 Comparative Results

We compared our system against heuristic (Round Robin), standard RL (DQN), and federated learning-based optimizers.

Table 2. Performance Comparison across Optimizers

Optimizer Type	**Latency Reduction (%)**	**Energy Saving (%)**	**QoS Score**
Round Robin	0	0	0.65
DQN	22	17	0.78
Federated Learning	30	21	0.82
Proposed System	**34**	**28**	**0.88**

6. Limitations and Future Work

6.1 Limitations

Despite promising results, our system has limitations. The federated learning model assumes stable network connections, which may not hold in highly mobile or volatile edge environments. Model convergence speed is also a concern when scaling to thousands of nodes. Furthermore, our RL-based models require retraining when faced with novel workload types.

Ethically, privacy concerns arise from transmitting metadata across tiers, although no user-identifiable data is included. Data protection protocols aligned with GDPR were implemented, but federated learning still presents risks if adversarial attacks are launched.

6.2 Future Directions

Future work will explore hybrid AI models combining reinforcement learning with game theory and agent-based simulation. We also plan to integrate anomaly detection mechanisms to automatically adapt model parameters when concept drift is detected in workload patterns.

Another promising avenue is the use of neuromorphic hardware for edge agents to enable ultra-low power inference in real-time. Expanding the testbed to real-world deployments in smart healthcare and transportation systems is also planned.

7. Conclusion

Multilayered AI workload optimizers represent a pivotal step toward efficient, adaptive, and scalable cross-platform data center architectures. This study presented an AI-driven framework capable of dynamic resource allocation using federated and reinforcement learning models. Through predictive modeling and real-time decision-making, the proposed system significantly improves latency, energy efficiency, and overall quality of service. As demands on data centers continue to evolve, such intelligent systems will be integral to the future of computing infrastructure.

References

1. Beloglazov, Anton, Jemal Abawajy, and Rajkumar Buyya. "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing." *Future Generation Computer Systems*, vol. 28, no. 5, 2012, pp. 755–768.

2. Sankaranarayanan S. (2025). From Startups to Scale-ups: The Critical Role of IPR in India’s Entrepreneurial Journey. International Journal of Intellectual Property Rights (IJIPR), 15(1), 1-24

3. Pulivarthy, P. (2022). Performance tuning: AI analyse historical performance data, identify patterns, and predict future resource needs. International Journal of Innovations in Applied Sciences and Engineering, 8(1), 139–155.

4. "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%." *DeepMind Blog*, 2016. Accessed 18 Apr. 2025.

5. Raza, Abid, Mohd Fadzil Hassan Othman, Sajjad A. Madani, and Sherali U. Khan. "Resource management in cloud computing: Taxonomy, prospects, and challenges." *Computers & Electrical Engineering*, vol. 47, 2015, pp. 186–203.

6. Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." *Nature*, vol. 518, no. 7540, 2015, pp. 529–533.

7. Sankaranarayanan S. (2025). Optimizing Safety Stock in Supply Chain Management Using Deep Learning in R: A Data-Driven Approach to Mitigating Uncertainty. International Journal of Supply Chain Management (IJSCM), 2(1), 7-22 doi: https://doi.org/10.34218/IJSCM_02_01_002

8. Pulivarthy, P., & Whig, P. (2025). Bias and fairness: Addressing discrimination in AI systems. In Ethical dimensions of AI development (pp.103-126). IGI Global. https://doi.org/10.4018/979-8-3693-4147-6.ch005

9. Calheiros, Rodrigo N., Rajiv Ranjan, Anton Beloglazov, César A. F. De Rose, and Rajkumar Buyya. "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms." *Software: Practice and Experience*, vol. 41, no. 1, 2011, pp. 23–50.

10. Mao, Hongzi, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. "Resource management with deep reinforcement learning." *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, 2016, pp. 50–56.

11. Mukesh, V. (2024). A Comprehensive Review of Advanced Machine Learning Techniques for Enhancing Cybersecurity in Blockchain Networks. ISCSITR-International Journal of Artificial Intelligence, 5(1), 1–6.

12. Chen, Tao, et al. "Edge cognitive computing based smart healthcare system." *Future Generation Computer Systems*, vol. 86, 2018, pp. 403–411.

13. Sankaranarayanan S. (2025). From Startups to Scale-ups: The Critical Role of IPR in India's Entrepreneurial Journey. International Journal of Intellectual Property Rights (IJIPR), 15(1), 1-24. doi: https://doi.org/10.34218/IJIPR_15_01_001

14. Pulivarthy, P. (2024). Gen AI Impact on the Database Industry Innovations. International Journal of Advances in Engineering Research (IJAER), 28(III), 1–10.

15. Mukesh, V., Joel, D., Balaji, V. M., Tamilpriyan, R., & Yogesh Pandian, S. (2024). Data management and creation of routes for automated vehicles in smart city. International Journal of Computer Engineering and Technology (IJCET), 15(36), 2119–2150. doi: https://doi.org/10.5281/zenodo.14993009

16. Sculley, D., et al. "Hidden technical debt in machine learning systems." *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, pp. 2503–2511.

17. Shi, Weisong, et al. "Edge computing: Vision and challenges." *IEEE Internet of Things Journal*, vol. 3, no. 5, 2016, pp. 637–646.

18. Mukesh, V. (2025). Architecting intelligent systems with integration technologies to enable seamless automation in distributed cloud environments. International Journal of Advanced Research in Cloud Computing (IJARCC), 6(1),5-10.

19. Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: State-of-the-art and research challenges." *Journal of Internet Services and Applications*, vol. 1, no. 1, 2010, pp. 7–18.

20. Li, Yiming, et al. "Survey of federated learning: Concepts, algorithms, and applications." *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, 2023, pp. 2361–2380.

21. Pulivarthy, P. (2024). Optimizing Large Scale Distributed Data Systems Using Intelligent Load Balancing Algorithms. AVE Trends in Intelligent Computing Systems, 1(4), 219–230.

22. Varghese, Blesson, and Rajkumar Buyya. "Next generation cloud computing: New trends and research directions." *Future Generation Computer Systems*, vol. 79, 2018, pp. 849–861.