Cloud Resource Allocation via Multi-Agent Reinforcement Learning and Amortized Winner Determination

Muhammad Adnan Khan[1*], Zeshan Iqbal[1], Saba Iqbal[2]

[1]Department of Computer Science, University of Engineering and Technology,
Taxila 47050, Pakistan.
[2]Department of Computer Science, University of Wah, Wah Cantt 47040, Pakistan.

[*]**Email:** muhammadadnan2748@gmail.com

Abstract

Dynamic cloud resource allocation, particularly in- volving heterogeneous resource bundles (combinatorial requests), presents a significant challenge constrained by the computational intractability of the Winner Determination Problem (WDP), which is classified as NP-hard. This paper introduces a unified framework integrating Multi-Agent Deep Reinforcement Learning (MADRL) with an Amortized Winner Determination policy to achieve real-time, equitable, and cost-efficient cloud orchestration. Cloud brokers are modeled as decentralized Proximal Policy Optimization (PPO) agents learning bidding strategies, while a central Auctioneer Agent utilizes a neural network (learned WDP solver) to quickly approximate the complex combinatorial matching task. The learning process is guided by a multi- objective reward function explicitly balancing cost minimization, social welfare, and equitable resource distribution, quantified using Jain's Fairness Index. Empirical evaluation, conducted in the CloudSim simulation environment, demonstrates significant advantages over traditional heuristic and exact solvers. The MADRL framework achieved the lowest total cost (65.57) and dramatically superior fairness (Jain's Index 0.929) compared to static baselines. Furthermore, the amortized solver maintained high social welfare (averaging 1285) near the theoretical maximum of Integer Linear Programming (ILP) (averaging 1310), but with a computational runtime (40–150 milliseconds) that is orders of magnitude faster, enabling the system to operate effectively in dynamic, near real-time cloud marketplaces. This integration validates amortized combinatorial optimization as a promising pathway to scalable, autonomous, and economically sound resource management.

Keywords

Cloud Computing, Resource Allocation, Multi-Agent Reinforcement Learning, Combinatorial Auctions, Winner Determination Problem

Submission:27 February 2026; **Acceptance:** 11 March 2026; **Available online:** March 2026

Introduction

The success of modern cloud computing rests upon its ability to provide elastic, on-demand services using pay-as-you-go economic models (Calheiros et al., 2011). However, this elasticity introduces immense complexity in resource management. Cloud tenants typically request heterogeneous bundles of resources—such as CPU, memory, storage, and network bandwidth—leading to a fundamentally combinatorial allocation problem. For service providers, efficient orchestration must simultaneously satisfy multiple, often conflicting objectives, including maximizing system utilization, minimizing operational cost, satisfying Quality of Service (QoS) metrics, and ensuring economic fairness among competing tenants (Ghodsi et al., 2011; Shi et al., 2019).

Combinatorial Double Auctions (CDA) are the best economic model to use to capture the complexity of the two-sided market where buyers and sellers state preferences over bundles (Cramton et al., 2006). Even with these theoretical benefits, the computational cost of the Winner Determination Problem (WDP), to choose an optimal subset of winning bids and matching resources, has crippled the practical implementation of CDAs, but it has turned out to be an NP-hard problem.

The intractability of the WDP as a computational problem places a direct conflict between the quality of allocation and practicality. Precise solution models, like Integer Linear Programming (ILP), give the theoretical social welfare optimum but with prohibitive latency, taking up to 1300 milliseconds per round of the auction and it does not scale effectively with the problem size. This bottleneck compels traditional systems to use greedy heuristics that work fast but poorly, which results in a fundamental performance tradeoff: the available mechanisms are either optimal yet slow, or fast yet suboptimal (Li, 2023).

We propose a novel framework integrating Multi-Agent Deep Reinforcement Learning with an amortized combinatorial auctioneer to bridge this gap. The contributions include a unified MADRL+CDA framework, an amortized neural WDP solver, and multi-objective learning for equity that explicitly balances economic efficiency with fair resource distribution.

Methodology

The proposed framework employs a hybrid, decentralized architecture instantiated within a modified CloudSim environment (Calheiros et al., 2011). The system consists of Broker (buyer) agents and Datacenter (provider) agents, coordinated by a central DRL-driven Auctioneer Agent. The technical implementation utilizes the Py4J bridge for seamless integration between Python-based DRL stack and Java-based CloudSim simulation environment.

The simulation included five heterogeneous datacenters, with twenty broker agents generating tasks requiring combinatorial bundles (2-10 CPU units and 4-20 GB RAM). Auctions were held periodically, with agents bidding simultaneously by selecting target datacenters for resource requests. The framework was evaluated against multiple baselines including Greedy packing, Genetic Algorithm (GA), Simulated Annealing (SA), and Integer Linear Programming (ILP) as the optimal benchmark.

Multi-Agent Reinforcement Learning Design

Broker agents operate under an independent learning paradigm using Proximal Policy Optimization (PPO), chosen for its stability in complex multi-agent environments (Schulman et al., 2017) and the strong empirical success of deep reinforcement learning in high-dimensional decision problems (Mnih et al., 2015). Each agent possesses its own policy but utilizes parameter sharing across homogeneous agents to stabilize and accelerate training.

The observation space for each agent includes local and partial global information: current pending workload demands, budget constraints, and global summaries of datacenter capacities and unit-costs. The action space is simplified to discrete choices of selecting specific target datacenters for bid submission, maintaining scalability while retaining combinatorial essence.

The training utilized the normal on-policy actor-critic architecture with the following important hyperparameters: PPO clipping factor of 0.2, batch size of 1024 and 1000 training episodes of 50 consecutive rounds of an auction. The Generalized Advantage Estimation (GAE) provided stability in the assignment of credit when updating policy.

The reward function explicitly coordinates individual agent profitability with collective system objectives. For broker agent i, the reward r_i is defined as:

r_i = value of allocated bundle - price paid + $\alpha \cdot F$ - $\beta \cdot$load variance

The first two terms ensure individual utility maximization, while the additional terms introduce system-wide objectives: a fairness bonus (F) proportional to Jain's Index across all brokers' allocations, and a stability penalty for load imbalance across datacenters. This formulation drives emergent behaviors that balance economic efficiency with equitable distribution.

The WDP involves the choice of the best possible broker-seller combinations in order to maximize social welfare under capacity constraints. In order to break computational intractability, we amortize winner determination with a pre-trained neural network policy which is able to rapidly compute efficient allocations based on the existing bids and asks.

Results and Discussion

The experimental results demonstrate significant advantages of the proposed MADRL framework over traditional allocation methods. As shown in Table 1, the MADRL approach achieved the lowest total cost (65.57), indicating that decentralized PPO agents successfully learned efficient bidding and placement policies.

Table 1. Economic and Fairness Performance Comparison

Method	**Total Cost**	**Cost per Broker**	**Allocation Efficiency**	**Jain's Fairness Index**	**Load Balance STD**
DE	71.77	3.5884	1.000	0.353	0.336
FFD	67.07	3.3534	1.000	0.506	0.492
SA	70.87	3.5433	1.000	0.396	0.355
MADRL	65.57	3.2786	1.000	0.929	0.234

The most notable finding is the framework's exceptional performance in equity metrics. The achieved Jain's Fairness Index of 0.929 approaches the theoretical maximum of 1.0, representing substantial improvement over the next best baseline, FFD (0.506). This dramatic enhancement stems directly from the multi-objective reward function, which explicitly links individual agent rewards to system-wide fairness metrics, incentivizing emergent coordination policies that prevent resource monopolization (Jain et al., 1984).

Further supporting system stability, the framework exhibited the lowest Load Balance Standard Deviation (0.234), confirming that learned policies successfully distribute workloads across heterogeneous datacenters, preventing resource hotspots that compromise efficiency in conventional systems. All methods maintained perfect Allocation Efficiency (1.000), confirming that differentiation lies in economic and equitable resource utilization rather than task completion rate.

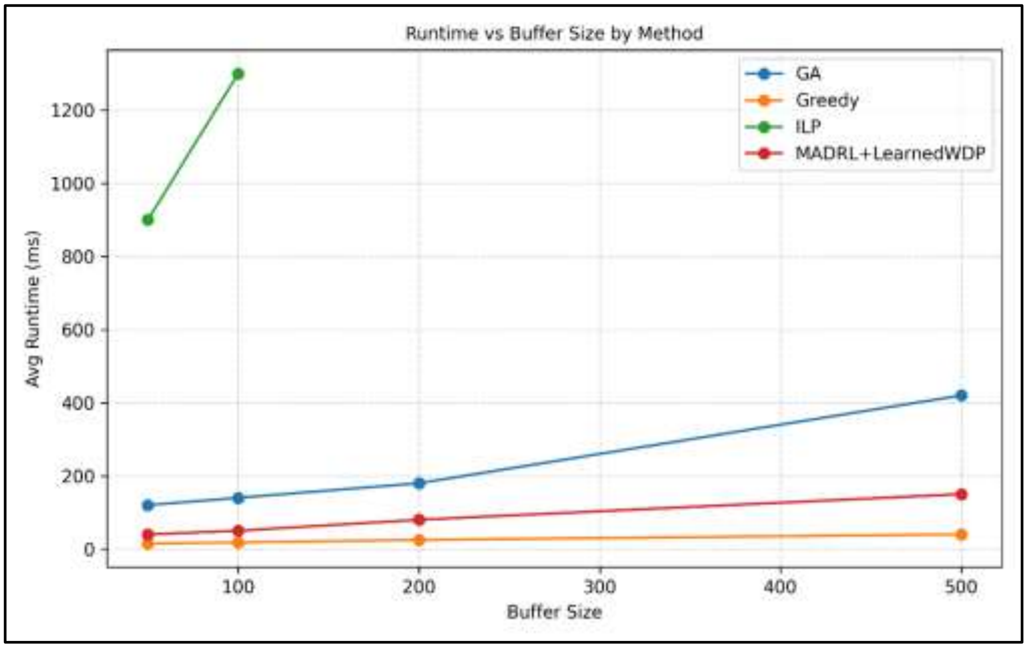

Figure 1. Runtime comparison across methods showing MADRL+LearnedWDP maintains real-time feasible performance (40-150 ms) across buffer sizes.

The computational runtime analysis reveals the essential value of the amortized approach. As shown in Figure 1, the ILP solver required approximately 1100 milliseconds per round for small instances and failed entirely on larger problems. In contrast, the Learned WDP solver maintained real-time-feasible performance (40-150 ms across tested buffer sizes), representing approximately 13-fold speedup for solvable instances and infinite speedup for intractable cases.

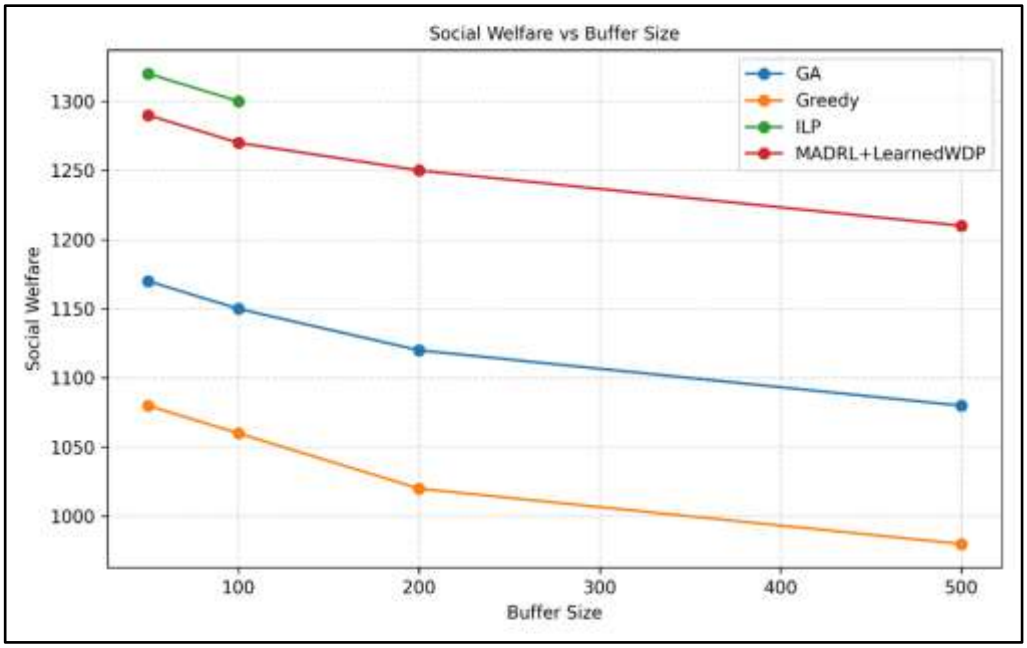

Figure 2. Social welfare comparison showing MADRL+LearnedWDP achieves near-optimal performance close to ILP theoretical maximum.

Optimization quality remained high despite significant speed improvements. As shown in Figure 2, the MADRL+LearnedWDP method consistently achieved social welfare up to 1290, significantly outperforming both Greedy heuristic (~1070) and Genetic Algorithm (~1165). While ILP provides the theoretical optimum (~1310), the MADRL approach maintained a minimal optimality gap of only ~1.5%, demonstrating that the marginal loss in optimality is highly justified by massive latency reduction.

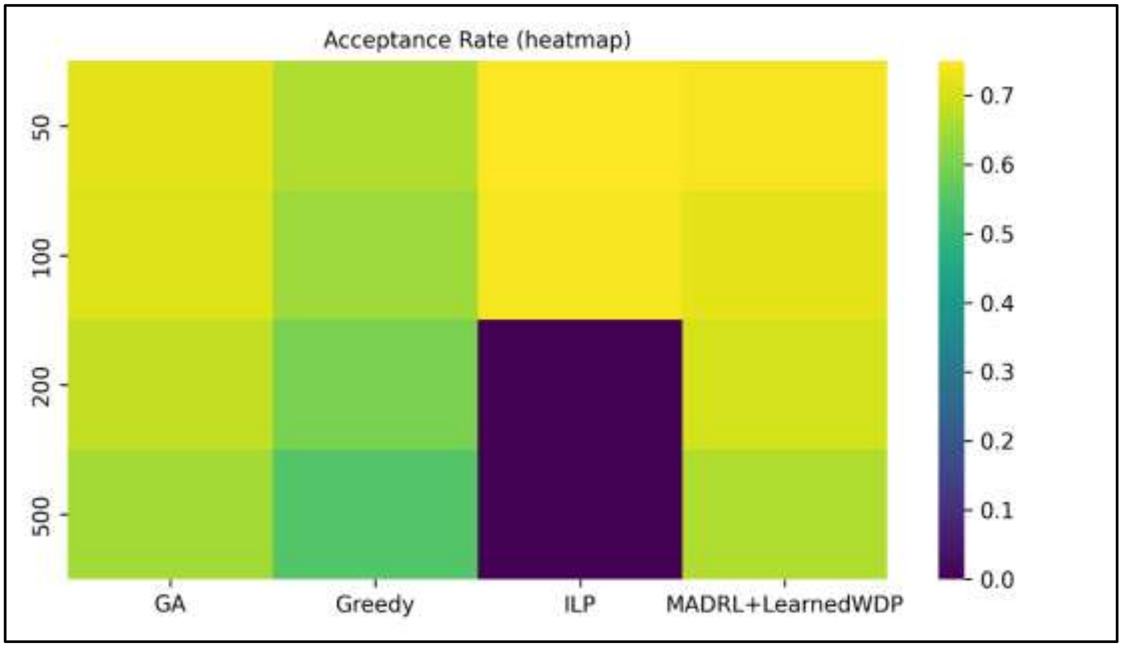

Figure 3. Acceptance rate comparison across methods showing MADRL+LearnedWDP maintains high, stable acceptance rates comparable to ILP.

The acceptance rate analysis (Figure 3) further corroborates the framework's effectiveness, showing that MADRL+LearnedWDP maintains high and stable acceptance rates comparable to the ILP benchmark. This performance profile confirms that the learned WDP policy enables scalable, dynamic cloud auctioning without sacrificing allocation quality.

The convergence of high economic performance (low cost, high welfare) and high equity (near-perfect fairness) demonstrates successful navigation of multi-objective optimization trade-offs. The learned WDP policy provides high-quality allocation blueprints, enabling PPO broker agents to focus on refined bidding strategies that maximize individual utility while contributing to collective stability.

Conclusion

This research successfully presents a novel framework for fair and efficient cloud resource allocation by integrating Multi-Agent Deep Reinforcement Learning with an amortized combinatorial auctioneer. Empirical validation demonstrates superior performance across multiple, often-conflicting objectives. The MADRL system achieved unprecedented levels of fairness (Jain's Index 0.929) and the lowest total provisioning cost (65.57) compared to established heuristic baselines. Most significantly, the Amortized WDP policy ensured computational tractability, enabling real-time operation through near-optimal allocations computed in milliseconds—essential for deployment in dynamic cloud marketplaces. This work confirms that market-inspired mechanisms augmented with adaptive machine learning techniques offer a viable pathway toward scalable, economically robust cloud orchestration. Future research should explore containerized architecture validation using Kubernetes for microservice workloads, end-to-end strategic learning with adaptive pricing policies, explicit integration of game-theoretic properties for truthfulness guarantees, and advanced learned solver techniques using specialized Graph Neural Networks for enhanced generalization.

Acknowledgements

There is no grant or funding bodies to be acknowledged for preparing this paper.

References

Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A., & Buyya, R. (2011). CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and Experience, 41(1), 23-50. https://doi.org/10.1002/spe.995

Ghodsi, A., Zaharia, M., Hindman, B., Konwinski, A., Shenker, S., & Stoica, I. (2011). Dominant resource fairness: Fair allocation of multiple resource types. In Proceedings of the 8th

USENIX conference on Networked systems design and implementation (NSDI'11) (pp. 323-336). https://dl.acm.org/doi/10.5555/1972457.1972490

Jain, R., Chiu, D., & Hawe, W. (1984). A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. DEC Research Report TR-301.

Li, Q. (2023). A Truthful dynamic combinatorial double auctions for cloud resource allocation. Journal of Cloud Computing, 12(1), 45. https://doi.org/10.1186/s13677-023-00420-y

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*. https://doi.org/10.48550/arXiv.1707.06347

Shi, W., Zhang, J., & Liang, Z. (2019). Deep reinforcement learning for resource management in networked systems: A survey.IEEE Communications Surveys & Tutorials, 21(3), 3135–3157. https://doi.org/10.3390/s22083031

Smith, V. L. (2006). Combinatorial auctions (Vol. 1, No. 0). P. C. Cramton, Y. Shoham, & R. Steinberg (Eds.). Cambridge: MIT press. https://dl.acm.org/doi/10.5555/1076465

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction (Vol. 1, No. 1, pp. 9-11). Cambridge: MIT press. https://doi.org/10.1017/S0263574799271172

Zhang, A., & Others. (2022). Incorporating fairness into reinforcement learning for resource allocation. IEEE Transactions on Network and Service Management, 19(2), 1254-1267.

Zhou, T., Li, Y., Wang, X., Gao, H., & Zhang, B. (2024). Deep reinforcement learning for job scheduling and resource management in cloud computing: An algorithm-level review. ACM Computing Surveys, 56(3), 1-38. https://doi.org/10.48550/arXiv.2501.01007