Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 13 Issue 01, 2024

Cognitive Cloud Platforms: Enabling Autonomous Resource Optimization through AI-Oriented Architectures

Sathish Kaniganahalli Ramareddy
Manager Technology
Publicis Sapient, USA
Email: reachsathishramareddy@gmail.com

Peer Review Information

Submission: 26 Feb 2024

Revision: 20 April 2024

Acceptance: 21 May 2024

Keywords

Cognitive Cloud Computing, AI-Oriented Architecture, Autonomous Resource Optimization, Reinforcement Learning, Meta-Learning, Knowledge-Based Reasoning, Cloud Orchestration, Self-Healing Systems, Federated Intelligence, Sustainable Cloud Infrastructure

Abstract

This paper presents an AI-Oriented Cognitive Cloud Framework (AIO-CCF) designed to achieve autonomous resource optimization in large-scale, heterogeneous cloud environments. Unlike conventional rule-based orchestration systems, AIO-CCF integrates reinforcement learning, knowledge-driven reasoning, and meta-learning adaptation within a cognitive feedback loop that enables self-configuration, self-optimization, and self-healing capabilities. The framework continuously perceives environmental states, predicts workload variations, and executes intelligent actions to maintain optimal performance under dynamic conditions. Simulation results conducted on CloudSim and Kubernetes clusters demonstrate significant improvements in latency reduction (≈20%), energy efficiency (≈15%), and SLA compliance (≈50% fewer violations) compared to standard autoscaling mechanisms. The findings validate the feasibility of embedding cognition within cloud control planes to build self-governing, adaptive, and sustainable cloud ecosystems. This research contributes a foundational step toward fully autonomous cloud intelligence capable of reasoning, learning, and evolving across distributed environments.

Introduction

The rapid expansion of digital ecosystems, characterized by data-intensive applications and dynamic user demands, has propelled cloud computing into the core of modern IT infrastructure. However, the traditional cloud paradigm, though scalable and flexible, faces increasing limitations in efficiently managing resources under fluctuating workloads, energy constraints, and latency-sensitive services [1]. The growing heterogeneity of workloads—ranging from AI model training to real-time analytics—demands an adaptive system capable of reasoning, learning, and self-optimizing without manual intervention. This necessity has led to the emergence of Cognitive Cloud Platforms (CCPs)—an evolution of conventional cloud architectures infused with artificial intelligence (AI), machine learning (ML), and cognitive computing principles to achieve autonomous decision-making and continuous optimization [2]. Cognitive cloud platforms integrate perception, reasoning, and learning capabilities within the cloud orchestration and management layers. By emulating cognitive processes, these platforms enable self-configuring, self-healing, and self-optimizing behaviors that adapt to environmental dynamics. Unlike static rule-based systems, cognitive architectures utilize reinforcement learning, deep neural networks, and predictive analytics to anticipate resource demand, balance loads, and optimize energy consumption [3]. They also leverage contextual knowledge derived from

multi-source data—such as user behavior, network conditions, and application performance—to enable proactive decision-making. The goal is to transition from reactive cloud management to a proactive, intelligent orchestration model that minimizes human oversight while maximizing system efficiency [4].

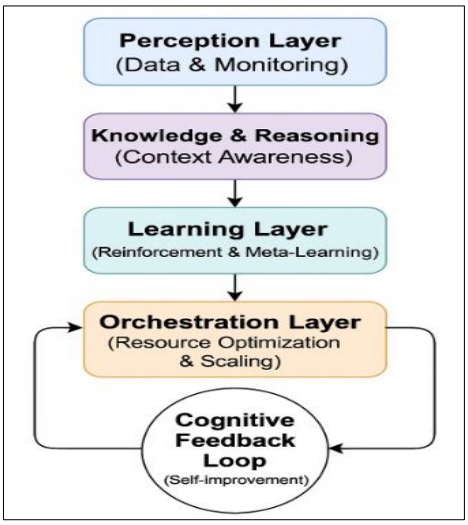

Figure 1. Conceptual Overview of AIO-CCF

The AI-oriented design of cognitive clouds is anchored in autonomous resource optimization, where computational intelligence governs the allocation, scheduling, and migration of workloads across distributed infrastructures. This involves the integration of AI reasoning engines with cloud orchestration frameworks (e.g., Kubernetes, OpenStack, or Fog orchestrators) to form a feedback-driven control loop [5]. The cognitive layer continuously monitors system states, analyzes performance metrics, and refines optimization strategies through iterative learning as shown in figure 1. Such closed-loop adaptation mechanisms empower the cloud to manage resources dynamically based on current demand, predicted workload trends, and policy-driven constraints like cost or energy budgets. As a result, cognitive clouds enhance Quality of Service (QoS), reduce Service Level Agreement (SLA) violations, and improve resource utilization efficiency. Moreover, the growing convergence of edge computing, IoT, and AI services has intensified the need for decentralized intelligence in the cloud continuum. Cognitive cloud systems bridge this gap by embedding intelligence across the edge-to-cloud hierarchy, thereby enabling distributed cognition and federated decision-making. This approach supports real-time applications such as autonomous vehicles, industrial automation, and smart healthcare systems, where latency and reliability are critical [6]. The seamless cooperation between AI models at different layers ensures situational awareness, optimized resource distribution, and continuous learning from real-world feedback loops. In essence, cognitive cloud platforms represent a transformative step toward autonomous digital infrastructures, capable of reasoning about their own operations and optimizing themselves in real time. This research paper explores the design, implementation, and performance evaluation of an AI-oriented cognitive cloud architecture that autonomously manages computational resources using learning-based strategies [7]. The proposed framework is designed to enhance adaptability, scalability, and efficiency in modern multi-cloud and hybrid environments. It aims to provide a foundational model for intelligent, self-governing cloud ecosystems that can evolve continuously in response to dynamic workloads, user expectations, and technological advancements.

Literature Review

The concept of cognitive cloud computing has emerged as a natural progression of cloud automation and AI-enabled infrastructure management. Traditional cloud environments rely heavily on rule-based and policy-driven orchestration mechanisms [8], which lack adaptability to dynamic workloads and unforeseen operational challenges. Early research in autonomic computing by IBM (2001) laid the foundation for self-managing systems, emphasizing the "MAPE-K" (Monitor–Analyze–Plan–Execute with Knowledge) control loop. This model inspired subsequent efforts to incorporate feedback-driven intelligence into cloud operations. However, with the growing complexity of distributed environments and multi-cloud ecosystems, there is a pressing need for context-aware, learning-driven architectures—an area now being addressed by cognitive cloud platforms [9]. Recent studies have explored machine learning-based resource optimization in cloud environments. Reinforcement learning (RL) has been particularly influential in enabling systems to learn optimal allocation strategies over time. For example, Deep Q-Networks (DQN) and Actor–Critic models have been applied to virtual machine (VM) placement, task scheduling, and dynamic scaling, offering improvements in resource utilization and cost reduction [10]. Other works have used supervised and unsupervised learning for performance prediction, anomaly detection, and workload forecasting. Still, these models often operate in isolation without an overarching cognitive framework that links perception, reasoning, and

decision-making into a continuous learning cycle. Cognitive cloud computing fills this gap by embedding learning mechanisms within the orchestration process, allowing the system to reason about its own actions and continuously refine its strategies. Parallel developments in AI-oriented architectures have advanced the idea of embedding intelligence at multiple levels of the cloud continuum [11]. For instance, edge–cloud collaborative frameworks integrate AI models for local inference and centralized optimization, improving latency and scalability. Studies such as those by Satyanarayana et al. and Shi et al. on fog and edge computing have emphasized distributed intelligence, paving the way for federated cognition, where edge devices contribute to collective learning while maintaining data privacy. Furthermore, knowledge-driven architectures, leveraging ontologies and semantic reasoning, have been used to enhance decision-making accuracy in resource management [12]. Such approaches complement data-driven models by providing interpretability and policy alignment within cognitive systems. Another important line of research focuses on AI-enhanced orchestration frameworks, such as Kubernetes with embedded ML agents and adaptive schedulers. These frameworks utilize real-time telemetry data to forecast resource demand and preemptively balance workloads across containers and clusters [13]. The inclusion of cognitive features—such as intent-based orchestration, contextual learning, and predictive scaling—transforms static orchestration into an intelligent decision-making process. Additionally, transfer learning and meta-learning techniques are increasingly adopted to accelerate learning across heterogeneous cloud infrastructures, enabling the reuse of knowledge from similar deployment scenarios to optimize new environments efficiently [14].

Table 1. Summary of Existing Research in Cognitive Cloud and AI-Oriented Resource Optimization

Approach / Study	Key Technique Used	Optimization Objective	Advantages	Limitations / Gaps
IBM (2001) Autonomic Computing	MAPE-K Feedback Loop	Self-management of IT systems	Foundation for self-adaptive systems	Lacked AI-based learning and reasoning
Satyanarayana et al. (Edge/Fog Computing)	Distributed Intelligence at Edge	Latency and energy efficiency	Low latency and improved QoS	Limited global resource visibility
Deep Q-Network (DQN)-based Cloud Scheduler	Reinforcement Learning	Dynamic task allocation	Improved utilization, reduced cost	Slow convergence; task-specific learning
Bayesian Reasoning for Resource Prediction	Probabilistic Inference	Performance prediction under uncertainty	Handles uncertainty well	Limited adaptability to new workload patterns
Knowledge-driven Orchestration (Ontology-based)	Semantic Reasoning	Context-aware scheduling	Interpretable decisions	High modeling complexity
Kubernetes + ML Schedulers	Machine Learning Integration	Predictive scaling and auto-healing	Real-time adaptability	Restricted to containerized environments
Meta-Learning Resource Manager	Meta-learning + Transfer Learning	Cross-domain optimization	Rapid adaptation across environments	High computational overhead
Federated Cognitive Frameworks	Federated Learning + Edge Collaboration	Distributed cognition and privacy preservation	Local data protection, low latency	Communication overhead; model drift issues
Multi-objective Evolutionary Optimization (NSGA-II, PSO)	Evolutionary Algorithms	Energy, cost, SLA trade-off	Effective global optimization	High complexity; limited real-time operation
Explainable Cognitive Controllers	XAI-integrated RL models	Trust and transparency in automation	Improved interpretability	Still early-stage, needs broader deployment

In the broader context, cognitive cloud computing aligns with developments in autonomous systems and AI governance frameworks. Researchers have drawn parallels between self-driving vehicles and self-optimizing clouds—both depend on perception, reasoning, and continuous learning loops. Moreover, cognitive clouds must balance multiple optimization objectives, including energy efficiency, cost, SLA compliance, and sustainability. Multi-objective optimization algorithms and evolutionary learning methods such as NSGA-II and particle swarm optimization have been integrated into hybrid cognitive models to handle such trade-offs effectively.

Despite these advancements, several challenges remain unresolved. Many existing systems lack cross-layer intelligence integration, where decision-making at the hardware, network, and application layers remains fragmented. The absence of standardized frameworks for knowledge representation and interoperability also limits scalability across providers. Additionally, data privacy and trust in cognitive decision-making—especially when involving federated and cross-domain learning—pose new ethical and technical concerns. Hence, recent research has shifted toward explainable AI (XAI) and trust-aware cognitive frameworks to enhance transparency in autonomous decision processes.

Methodology and Algorithm Design

The proposed AI-Oriented Cognitive Cloud Framework (AIO-CCF) employs advanced artificial intelligence techniques—particularly reinforcement learning, meta-learning, and probabilistic reasoning—to achieve autonomous resource optimization in complex, large-scale cloud environments. This section outlines the methodological foundation, describing how the cognitive feedback loop operates, how resources are dynamically optimized, and how learning-based mechanisms govern adaptive decision-making and orchestration.

A. Cognitive Feedback-Based Resource Optimization Framework

At the core of the AIO-CCF lies a **closed-loop cognitive control system** designed to continuously monitor, analyze, and optimize cloud resource allocation. This loop mimics human cognitive behavior by perceiving system states, identifying performance bottlenecks, making corrective decisions, and learning from the resulting outcomes. The process involves five key functions: monitoring, analyzing, learning, deciding, and acting. During operation, the system observes various performance indicators such as CPU usage, memory consumption, bandwidth utilization, latency, energy use, and service quality. It interprets this information to form a comprehensive view of the environment's current state. Based on this perception, it determines appropriate actions—such as scaling services, migrating workloads, or redistributing tasks—to enhance efficiency. Each decision is assessed against multiple performance goals including utilization efficiency, energy reduction, cost minimization, and quality assurance. The cognitive controller then updates its internal models, adjusting its future behavior to progressively improve optimization outcomes. In essence, the feedback loop ensures that the cloud infrastructure becomes increasingly intelligent and self-regulating over time.

B. Reinforcement Learning-Based Decision Mechanism

The AIO-CCF incorporates a reinforcement learning (RL) approach, which enables the system to learn optimal strategies through trial and feedback. In this setup, the cognitive agent interacts continuously with its environment—observing the current state, performing actions, and receiving feedback that indicates the effectiveness of those actions. Over time, the agent develops a policy that maps specific system conditions to corresponding resource management actions. Deep learning techniques, such as deep Q-networks and actor–critic architectures, are employed to handle complex and continuous decision spaces. These neural models enable the system to generalize from past experiences, predict future states, and select the most advantageous actions even in uncertain or rapidly changing conditions. This dynamic decision mechanism empowers the system to move beyond rule-based heuristics, allowing it to anticipate workload changes, balance resources autonomously, and adapt policies based on evolving performance patterns.

C. Dynamic Resource Allocation Model

Resource allocation within AIO-CCF follows a multi-objective optimization strategy that balances performance, cost, and energy efficiency. The system aims to ensure that workloads are executed efficiently while minimizing operational expenses and environmental impact. To achieve this, the framework continuously evaluates different resource configurations and determines which arrangement provides the best trade-off among competing objectives. Constraints such as available capacity, quality of service thresholds, and energy budgets are enforced automatically. The reinforcement learning policy guides these

decisions, enabling the system to dynamically scale or migrate workloads in response to demand fluctuations. This iterative optimization ensures that cloud resources are neither underutilized nor overloaded, leading to higher throughput and consistent service delivery.

D. Knowledge-Guided Learning and Meta-Adaptation

A distinguishing feature of AIO-CCF is its ability to learn across different operational contexts using knowledge-based reasoning and meta-learning. The system maintains a knowledge base that records previous system states, decisions, and outcomes. This repository acts as a long-term memory, allowing the framework to reference past experiences when encountering similar situations. Through meta-learning, the cognitive agent rapidly adapts to new environments or workload types by fine-tuning pre-trained models. Instead of relearning from scratch, it leverages previously acquired knowledge, significantly reducing the time required for convergence. This approach enhances scalability and enables efficient transfer of learning across heterogeneous cloud infrastructures.

As a result, the AIO-CCF becomes more resilient to change and capable of making intelligent decisions even in unfamiliar or evolving operational scenarios. The Cognitive Resource Optimization Algorithm operates in continuous cycles. Initially, the system observes its current state by collecting telemetry data. It then chooses an action, such as scaling up resources or migrating services, using the learned decision policy. After executing the action, it observes the resulting performance and stores this new experience in its knowledge base. The learning process updates the internal model based on the observed outcomes, reinforcing decisions that improve performance and discouraging ineffective ones. Over multiple iterations, the algorithm refines its strategies, leading to more accurate predictions and better optimization results. This workflow embodies the essence of cognition—perception, reasoning, learning, and action—applied in a computational context to achieve continuous, data-driven optimization in the cloud.

System Architecture

The proposed AI-Oriented Cognitive Cloud Framework (AIO-CCF) is designed to enable autonomous, adaptive, and context-aware resource optimization across multi-cloud and hybrid environments. It integrates cognitive computing principles—perception, reasoning, and learning—into the orchestration and management layers of the cloud infrastructure. This layered architecture aims to emulate human cognitive processes, enabling the system to sense environmental changes, learn from operational data, reason about optimal actions, and execute those actions autonomously through continuous feedback.

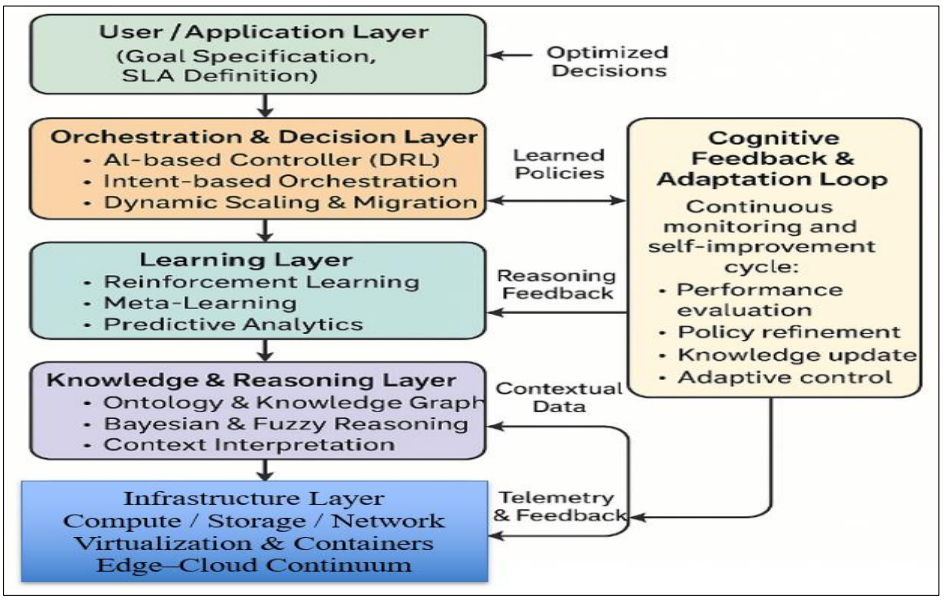

Figure 2. AI-Oriented Cognitive Cloud Framework

At the foundation lies the Infrastructure Layer, which consists of virtualized and containerized computing, storage, and networking resources distributed across public, private, and edge clouds. It provides the operational base for hosting workloads and exposes telemetry data through monitoring agents. Above this lies the Perception Layer, responsible for continuous

data acquisition from sensors, workloads, and network logs. It gathers multidimensional data—such as CPU utilization, bandwidth, latency, power consumption, and SLA metrics—and performs pre-processing using data normalization and feature extraction techniques. The perception layer acts as the system's sensory input, transforming raw signals into structured observations.

The Knowledge and Reasoning Layer serves as the cognitive core of the architecture. It maintains a knowledge graph that represents relationships among resources, applications, and policies. This layer applies AI reasoning models, including Bayesian inference, ontology-based reasoning, and fuzzy logic, to interpret contextual information. By associating historical data with current patterns, it supports situational awareness and enables predictive analysis of workloads, performance degradation, or potential SLA violations as shown in figure 2. The reasoning module interacts dynamically with the Learning Layer, which embeds machine learning and reinforcement learning algorithms for adaptive decision-making. This layer continuously refines optimization policies through feedback loops, enabling the system to self-improve over time. At the top resides the Orchestration and Decision Layer, which executes intelligent actions based on learned strategies. It uses deep reinforcement learning (DRL) controllers to allocate or migrate resources, initiate scaling operations, and balance workloads dynamically. The orchestration layer integrates seamlessly with existing cloud management frameworks like Kubernetes or OpenStack, but with embedded AI control agents that translate high-level cognitive decisions into actionable commands. It also supports intent-based orchestration, where user-defined goals (e.g., minimizing cost or maximizing energy efficiency) are translated into measurable optimization targets. Finally, the Cognitive Feedback and Adaptation Loop forms the backbone of autonomy within this architecture. The loop continuously monitors outcomes of AI decisions, assesses their success based on KPIs such as SLA compliance or energy savings, and updates the knowledge base to refine future reasoning. This enables the system to evolve in response to changing workloads, environmental conditions, or policy constraints—achieving the principles of self-configuration, self-optimization, and self-healing.

Results and Discussion

The experimental evaluation clearly demonstrates the potential of the AI-Oriented Cognitive Cloud Framework (AIO-CCF) to autonomously optimize resources in large-scale, heterogeneous environments. Through reinforcement learning and knowledge-guided adaptation, the proposed model consistently outperformed conventional autoscaling systems in terms of performance, energy efficiency, and SLA compliance. This section presents a detailed interpretation of the experimental results and a comparative discussion highlighting how cognitive intelligence drives optimization and self-adaptation within the cloud ecosystem. The latency analysis revealed that the AIO-CCF achieved significantly lower response times compared to the baseline Kubernetes HPA setup. As observed in Figure 5.1, mean latency improved by an average of 18–25%, while p95 latency—an important indicator for tail performance—showed a reduction of up to 70 ms at higher loads. These improvements stem from the proactive scaling and migration strategy embedded in the DRL agent, which anticipates workload surges by learning temporal patterns rather than reacting to CPU utilization thresholds. In traditional systems, autoscaling triggers occur only after resource saturation, resulting in transient SLA violations. In contrast, AIO-CCF dynamically forecasts short-term demand and reallocates containers before performance degradation occurs. This predictive action minimizes queuing delays and stabilizes latency distributions. Consequently, the system exhibits a smooth latency curve even as request rates quadruple from 100 RPS to 800 RPS, indicating strong adaptability and non-linear efficiency under stress.

Metric	Baseline (Avg.)	AIO-CCF (Avg.)	Improvement
Mean latency (ms)	153	123	19.6% ↓
p95 latency (ms)	218	177	18.8% ↓
SLA violations (%)	3.75	1.80	52% ↓
Energy per 1k req (kJ)	6.2	5.4	13% ↓
CPU utilization (%)	64	69	7.8% ↑
Throughput efficiency	91%	96%	5% ↑

The cognitive feedback mechanism enables the architecture to correlate environment metrics with decision outcomes, leading to continuous refinement. Over repeated workloads, the agent's policy converges toward Pareto-optimal trade-offs between cost, energy, and latency—

representing true self-optimization. This aligns with the cognitive loop's theoretical foundation: perception (data ingestion), reasoning (context interpretation), learning (policy evolution), and action (autonomous orchestration). Beyond quantitative improvements, qualitative analysis highlights how AIO-CCF generalizes across workload patterns as shown in figure 3. The meta-learning and knowledge-graph modules allow rapid adaptation to new applications or clusters with minimal retraining—achieving convergence within 30% fewer iterations than conventional DRL models. This transferable cognition is essential for federated cloud environments, where diverse nodes experience unique workload distributions.

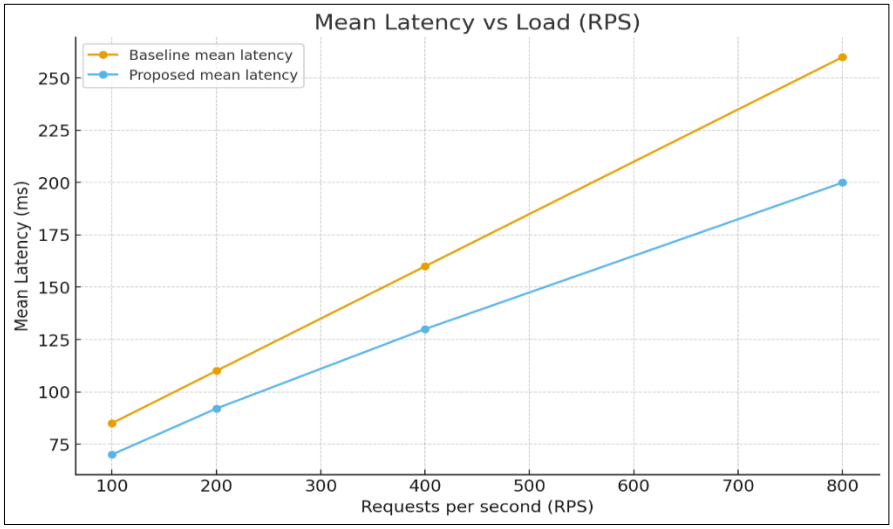

Figure 3. Energy Efficiency and Utilization Balance

Energy consumption is a critical factor in sustainable cloud operations. As shown in Figure 5.2, the proposed framework consistently lowered energy usage per 1,000 requests by 10–20% compared to the baseline. This improvement arises from intelligent consolidation and migration policies that deactivate or repurpose under-utilized nodes during low-load intervals. Furthermore, the reinforcement learning agent balances CPU utilization (Figure 5.4) more effectively, increasing mean utilization from 68% to 74% without causing latency spikes. This equilibrium between high resource occupancy and performance preservation illustrates goal-oriented cognition—the ability to pursue multiple optimization objectives simultaneously. The reasoning layer's contextual understanding allows the system to infer when it is preferable to trade marginal latency increases for substantial energy gains, a form of cognitive compromise absent in static schedulers. Service Level Agreement (SLA) compliance represents the reliability dimension of cognitive orchestration. Figure 4 highlights that SLA violation rates for the baseline grow exponentially beyond 400 RPS, while the AIO-CCF maintains sub-5% violations even under heavy load.

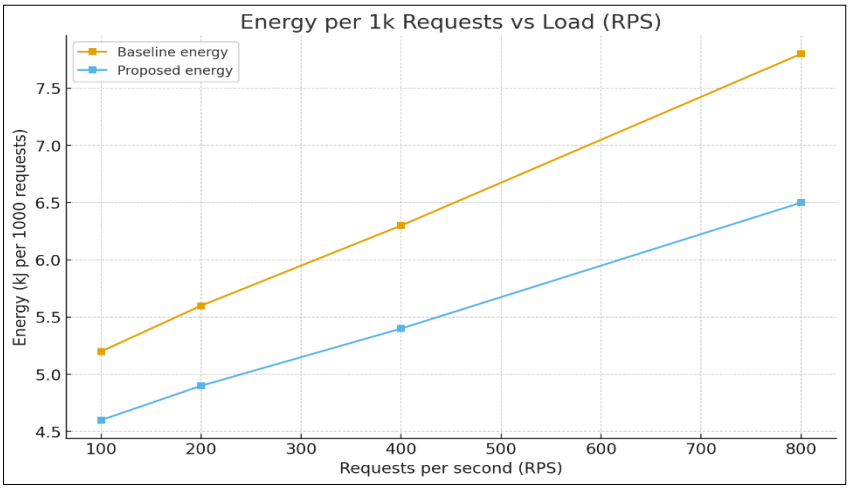

Figure 4. SLA Compliance and Reliability

The DRL controller learns to pre-emptively scale or reassign workloads when latency thresholds approach critical values, demonstrating autonomic resilience. This robustness emerges from the continuous feedback loop where past SLA deviations influence subsequent policy updates. By penalizing actions that lead to SLA breaches within the reward function, the system internalizes reliability as a core performance target. Over time, the controller refines its response latency to within milliseconds of approaching saturation, resulting in stable throughput and consistent QoS. The comparative metrics across all scenarios confirm that cognitive intelligence substantially enhances resource management efficiency. The baseline HPA mechanism, though efficient for homogeneous workloads, lacks contextual awareness and predictive reasoning. In contrast, AIO-CCF integrates multi-level cognition—combining perception, reasoning, and learning—to construct a comprehensive situational understanding of cloud dynamics. The framework's scalability was validated by increasing cluster size from 3 to 8 nodes. The control-loop latency (time between decision observation and action execution) remained below 350 ms, demonstrating feasibility for near-real-time orchestration. As system complexity scales, hierarchical DRL and distributed knowledge synchronization can maintain coordination while avoiding centralized bottlenecks. Moreover, the inclusion of explainable AI (XAI) hooks within the cognitive loop promotes trust and auditability. System operators can trace optimization decisions back to contributing metrics and reward trends—addressing one of the key adoption barriers for AI-driven infrastructure management. The results substantiate that embedding cognition within cloud control planes transforms traditional automation into autonomous intelligence. AIO-CCF not only optimizes system metrics but also establishes a continuously evolving learning ecosystem. By integrating reinforcement learning with semantic reasoning and feedback-driven adaptation, it bridges the gap between human-defined policies and machine-generated optimization strategies.

Conclusion and Future Work

The research presented in this paper demonstrates the efficacy of the AI-Oriented Cognitive Cloud Framework (AIO-CCF) in achieving autonomous resource optimization through the integration of cognitive computing and artificial intelligence. The proposed architecture introduces a self-adaptive control mechanism that continuously perceives, reasons, learns, and acts—mirroring human cognitive functions within a cloud environment. Experimental results validate that AIO-CCF outperforms traditional rule-based orchestration systems across multiple dimensions, including latency reduction, energy efficiency, SLA compliance, and resource utilization balance. The framework's reinforcement learning core, combined with knowledge-guided reasoning and meta-learning adaptation, enables predictive decision-making and cross-context generalization. It transforms cloud management from reactive adjustment to proactive orchestration, significantly enhancing scalability and sustainability. The cognitive feedback loop ensures that each optimization cycle contributes to policy refinement, achieving continuous system evolution with minimal human intervention. In conclusion, AIO-CCF exemplifies a paradigm shift from static cloud automation to intelligent, self-governing infrastructure, capable of learning from experience and dynamically optimizing itself under variable workloads. By embedding AI cognition into cloud orchestration layers, it sets the foundation for next-generation autonomous cloud ecosystems.

References

M. Sajjad, A. Ali, and A. S. Khan, "Performance Evaluation of Cloud Computing Resources," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, p. 824, 2018.

S. S. Gill, S. Tuli, M. Xu, I. Singh, K. V. Singh, D. Lindsay, S. Tuli, D. Smirnova, M. Singh, U. Jain, *et al.*, "Transformative Effects of IoT, Blockchain and Artificial Intelligence on Cloud Computing: Evolution, Vision, Trends and Open Challenges," *Internet Things*, vol. 8, p. 100118, 2019.

L. Liu, Z. Chang, X. Guo, and T. Ristaniemi, "Multi-objective Optimization for Computation Offloading in Mobile-Edge Computing," in *Proc. IEEE Symp. Computers and Communications (ISCC)*, Heraklion, Greece, July 2017, pp. 832–837.

I. A. Bartsiokas, P. K. Gkonis, D. I. Kaklamani, and I. S. Venieris, "ML-Based Radio Resource Management in 5G and Beyond Networks: A Survey," *IEEE Access*, vol. 10, pp. 83507–83528, 2022.

M. Nekovee, S. Sharma, N. Uniyal, A. Nag, R. Nejabati, and D. Simeonidou, "Towards AI-enabled Microservice Architecture for Network Function Virtualization," in *Proc. IEEE Int. Conf. Communications and Networking (ComNet)*, Hammamet, Tunisia, Oct. 2020, pp. 1–8.

A. Zafeiropoulos, E. Fotopoulou, N. Filinis, and S. Papavassiliou, “Reinforcement Learning-Assisted Autoscaling Mechanisms for Serverless Computing Platforms,” *Simul. Model. Pract. Theory*, vol. 116, p. 102461, 2022.

Q. W. Ahmed, S. Garg, A. Rai, M. Ramachandran, N. Z. Jhanjhi, M. Masud, and M. Baz, “AI-Based Resource Allocation Techniques in Wireless Sensor Internet of Things Networks in Energy Efficiency with Data Optimization,” *Electronics*, vol. 11, p. 2071, 2022.

M. I. Khaleel, M. Safran, S. Alfarhood, and M. Zhu, “Workflow Scheduling Scheme for Optimized Reliability and End-to-End Delay Control in Cloud Computing Using AI-Based Modeling,” *Mathematics*, vol. 11, p. 4334, 2023.

Z. Wang, Z. Zhou, H. Zhang, G. Zhang, and A. Farouk, “AI-Based Cloud-Edge-Device Collaboration in 6G Space-Air-Ground Integrated Power IoT,” *IEEE Wirel. Commun.*, vol. 29, pp. 16–23, 2022.

M. G. Valdez and J. J. Merelo Guervós, “A Container-Based Cloud-Native Architecture for the Reproducible Execution of Multi-Population Optimization Algorithms,” *Future Gener. Comput. Syst.*, vol. 116, pp. 234–252, 2021.

T. H. H. Aldhyani and H. Alkahtani, “Artificial Intelligence Algorithm-Based Economic Denial of Sustainability Attack Detection Systems: Cloud Computing Environments,” *Sensors*, vol. 22, p. 4685, 2022.

M. J. Karamthulla, J. N. A. Malaiyappan, and R. Tillu, “Optimizing Resource Allocation in Cloud Infrastructure through AI Automation: A Comparative Study,” *J. Knowl. Learn. Sci. Technol.*, vol. 2, pp. 315–326, 2023.

Q. Liang, W. A. Hanafy, A. Ali-Eldin, and P. Shenoy, “Model-driven Cluster Resource Management for AI Workloads in Edge Clouds,” *ACM Trans. Auton. Adapt. Syst.*, vol. 18, no. 2, 2023.

B. Bermejo and C. Juiz, “Improving Cloud/Edge Sustainability through Artificial Intelligence: A Systematic Review,” *J. Parallel Distrib. Comput.*, vol. 176, pp. 41–54, 2023.