Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 1, February 2026, pp. 470~483
ISSN: 2302-9285, DOI: 10.11591/eei.v15i1.10352

Resource cost management in cloud service environment

Kong FanYong, Fang-Fang Chua, Amy Hui-Lan Lim
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia

Article Info

Article history:

Received Apr 7, 2025
Revised Oct 3, 2025
Accepted Dec 6, 2025

Keywords:

Cloud cost
Cost optimization
Finops
Machine learning
Multi-cloud management
Reinforcement learning
Serverless computing

ABSTRACT

Cloud computing has revolutionized information technology (IT) infrastructure by enabling on-demand access to scalable resources. However, the elasticity and complexity of cloud billing models introduce significant challenges for effective resource cost management. This paper proposes a hybrid framework integrating statistical models auto regressive integrated moving average (ARIMA), machine learning techniques long short-term memory (LSTM), and optimization methods deep deterministic policy gradient (DDPG) to forecast and manage cloud costs with enhanced accuracy and adaptability. The framework is empirically validated using synthetic billing datasets and real-world cloud provider data, with performance evaluated via root mean square error (RMSE) and mean absolute percentage error (MAPE) metrics. Results demonstrate 15-25% improvement in cost prediction accuracy over baseline models and up to 20% cost savings through dynamic resource allocation. The framework extends beyond traditional VM-based workloads to support serverless computing (amazon web services (AWS) Lambda and Azure Functions) and container-based applications (Docker and Kubernetes), addressing the growing adoption of microservices architectures. Comparative analysis with existing tools (AWS Cost Explorer and Azure Advisor) reveals superior adaptability in multi-cloud environments. The paper concludes with discussions of emerging paradigms including FinOps practices, AIOps automation, and sustainability-aware resource allocation, outlining future research directions toward explainable AI-driven cost governance.

Corresponding Author:

Fang-Fang Chua
Faculty of Computing and Informatics, Multimedia University
Cyberjaya, Selangor, Malaysia
Email: ffchua@mmu.edu.my

1. INTRODUCTION

The pandemic-induced shift to cloud computing has fundamentally transformed enterprise information technology (IT) infrastructure deployment and management. Cloud services enable organizations to obtain computing, storage, and network resources on demand through service-oriented models, significantly reducing dependency on upfront capital expenditures [1]. This shift has increased business agility and innovation while reducing operational costs and risks associated with maintaining physical infrastructure. Global cloud adoption continues to grow, with organizations across all sectors migrating critical workloads to cloud environments [2]. According to Gartner forecasts, worldwide public cloud end-user spending reached $494.7 billion in 2022 and is projected to exceed $700 billion by 2025, reflecting the accelerating pace of cloud adoption across industries [2].

However, the same characteristics that make cloud services attractive—elasticity, on-demand provisioning, and pay-as-you-go billing—create significant complexity for resource cost management. As cloud deployments expand, enterprises face mounting cost pressures from increasing resource scale and diverse billing models [3]. In dynamic cloud environments where resources can rapidly scale based on workload requirements, effective cost management becomes particularly challenging. Recent surveys indicate that financial governance is now the highest-ranked priority for cloud adopters [4], with cost control consistently ranking among the top three challenges facing chief information officers (CIOs) and IT leaders.

Organizations waste approximately 30-35% of their cloud spend due to suboptimal resource allocation practices [5], highlighting the urgent need for advanced cost optimization approaches. Existing literature primarily focuses on individual techniques such as time series forecasting or reinforcement learning but lacks integrated frameworks combining statistical, machine learning, and optimization methods for holistic cost management [6], [7]. Key gaps include insufficient real-time adaptability to multi-cloud setups, limited interpretability in predictions, and inadequate consideration of trade-offs between cost, performance, security, and compliance [8], [9]. Reinforcement learning approaches have shown particular promise in dynamic resource allocation scenarios, demonstrating adaptability to changing workload patterns through continuous environmental interaction [10]. Deep learning methods for workload prediction have proven effective in capturing temporal dependencies in cloud resource utilization patterns [11].

This paper proposes a novel hybrid cloud resource cost management framework that integrates statistical models (auto regressive integrated moving average (ARIMA) with parameters p=1, d=1, and q=1), machine learning techniques (long short-term memory (LSTM) with 50 units), and optimization methods (deep deterministic policy gradient (DDPG) reinforcement learning) to predict resource usage, forecast costs, and recommend optimal allocations in multi-cloud environments. Our contributions include: i) a unified architecture for multi-cloud environments (amazon web services (AWS), Azure, and google cloud platform (GCP)) enhancing cost prediction accuracy by 15-25% over baseline models, ii) incorporation of real-time optimization with considerations for security and compliance risks, iii) empirical validation through simulations using synthetic and public datasets demonstrating up to 20% cost savings, and iv) extension beyond traditional VM-based workloads to support serverless computing environments and container-based applications. Table 1 summarizes the comparative performance of our approach against existing methods. This framework addresses identified gaps by providing interpretable, adaptive cost management suitable for hybrid and multi-cloud deployments.

Table 1. Performance comparison of cost management approaches

Model	RMSE	MAPE (%)	Cost savings (%)	Adaptability
ARIMA	12.5	18.2	8.4	Low
LSTMs	9.8	14.5	12.1	Medium
Hybrid	7.8	7.6	20.5	High
Serverless	6.9	6.9	23.2	High
Container	7.1	7.1	21.8	High

The remainder of this paper is organized as follows: section 2 reviews relevant literature on quota optimization, cost forecasting, pricing strategies, and cost observability. Section 3 describes the methodologies, datasets, evaluation metrics, and presents our proposed hybrid framework with detailed architecture. Section 4 provides empirical simulation results, comparative analysis, architectural considerations, and future research directions. Section 5 concludes with key findings and contributions.

2. LITERATURE REVIEW

2.1. Quota optimization

Cloud resource quota optimization has emerged as a critical component of effective cloud cost management, addressing the challenge of balancing resource availability with cost efficiency. Traditional cost management strategies designed for static on-premises infrastructure prove inadequate in dynamic cloud environments [6]. This review provides a comprehensive view of the evolving landscape of cloud resource cost management, from rule-based systems to advanced machine learning methodologies.

The review critically analyzes existing techniques for resource quota optimization, cost forecasting, and pricing strategies in cloud settings, outlining key challenges and limitations of current methodologies. Organizations waste approximately 30-35% of cloud spend due to inefficient resource allocation [5], highlighting the urgent need for sophisticated optimization approaches. These inefficiencies stem from over-provisioning to meet peak demands, underutilization during off-peak periods, and lack of visibility into resource consumption patterns across distributed teams.

Recent work emphasizes deep learning, multi-objective optimization, and causal reasoning approaches as particularly relevant for cloud cost management [7]. This study provides a systematic framework to understand the intricate relationships between resource allocation, utilization, and cost in cloud environments over time. The novelty lies in comprehensively benefiting both academic researchers and industrial practitioners working through challenges of resource provisioning and optimization. For researchers, it identifies promising avenues for further inquiry and knowledge gaps. For practitioners, it presents practical guidance on adopting appropriate cost management strategies based on organizational requirements and cloud deployment characteristics. The analysis covers both technical and organizational aspects since effective solutions must address algorithmic challenges and human implementation dynamics [7].

2.2. Cost forecasting

Accurate cloud resource cost prediction is essential for efficient budget planning and resource management. Many cloud cost forecasting techniques rely on time series analysis. Reinforcement learning-based autoscaling approaches have demonstrated significant potential for dynamic resource provisioning, learning optimal policies through continuous interaction with cloud environments and adapting to varying workload patterns [10].

Ruan *et al.* [12] introduced a hybrid model for cloud cost prediction combining time series and machine learning. They integrate classical time series models like ARIMA with neural networks to capture both linear and nonlinear trends in cost data. Results showed their ensemble model achieved 23% reduction in mean absolute percentage error (MAPE) compared to single models. Deep learning approaches for workload prediction in storage systems have shown remarkable effectiveness in capturing temporal dependencies, particularly for complex cloud resource utilization patterns exhibiting non-linear characteristics [12]. Guzek *et al.* [13] used seasonal decomposition methods to identify repeatable patterns in cloud usage, assisting in predicting workloads with weekly and monthly seasonality components.

Graph neural networks (GNNs) offer powerful approaches for modeling complex dependencies between cloud resources and costs. GNNs represent cloud resources as nodes and relationships as edges, capturing intricate interaction patterns governing cost dynamics. However, GNNs struggle with interpretability, limiting their use in environments requiring cost forecasting explainability [14]. Other researchers have investigated combining causal reasoning with deep learning to enrich cost prediction models with interpretability and robustness. These approaches aim to explain predictions—beyond predictive accuracy—by constructing causal graph models between resource usage and cost factors. Knowledge distillation techniques that approximate complex model behavior with more interpretable shallow models represent another promising direction for addressing the cost forecasting “black box problem” [15].

Multi-cloud [16] deployments introduce additional complexity to cost prediction due to heterogeneous pricing models and billing structures. Anh [17] addresses cross-cloud cost measurement and unified management problems, suggesting a common data model to normalize cost metrics between platforms. Their plugin-based solution encourages consistency in collecting and analyzing cost data from various cloud providers, though plugins require continual maintenance as cloud services evolve.

Real-time cost monitoring capabilities have significantly improved. Real-time cost monitor designed by [18] combines functionalities of data acquisition, processing, visualization, and alerting to deliver timely insights into cloud spending. These systems excel at tracking current costs but typically lack predictive capabilities to mitigate upcoming cost increases. Transfer learning methods have been applied to cost forecasting, showing promise for solving cold start problems when limited historical data is available. Transfer learning approaches improve prediction accuracy even for new cloud deployments or services with minimal historical data by leveraging knowledge from similar tasks or domains [19].

2.3. Pricing strategies

Pricing plays a key role in cloud cost management for maximizing provider revenue while minimizing consumer expenses. This section reviews significant studies on pricing models [20] for cloud resource economics and performance. Based on market mechanisms, [21] proposed a market-based pricing strategy where resource prices adjust according to supply and demand. This method determines benchmark prices through fixed and variable costs paired with market dynamics, providing real-time tuning to direct user behavior. Their model enabled personalized pricing that increased total revenue by 15-20% while ensuring customer satisfaction by analyzing price sensitivity across client segments.

Hybrid pricing models combining multiple pricing techniques have generated significant interest due to their potential to serve unique user needs. Introduced a pricing model effectively combining reserved instances, on-demand instances, and spot instances to address different usage patterns and cost-sensitivity requirements [22]. They propose a multi-objective optimization model that simultaneously maximizes provider revenue and minimizes user costs while considering competition, budget constraints, and quality of service.

The study demonstrates how combining appropriate pricing models achieves substantially better revenue than conventional fixed-price models while generating better overall user satisfaction through choice. However, implementation challenges include complexity of optimal allocation across different pricing models, effects on billing systems, and necessity for clear guidance to users to avoid inappropriate pricing. Auction-based and negotiation-based pricing mechanisms represent another active research area, enabling dynamic price discovery according to actual market conditions rather than fixed schedules. Although potentially more optimal in matching supply to demand, these methods introduce additional complexity and uncertainty that may not be appreciated by budget-constrained users [23]. Sharghivand *et al.* [24] offered guaranteed performance bounds in auction models designed with uncertainty limitation while maintaining economic efficiency.

Machine learning techniques have been applied to pricing optimization with promising results in predicting price elasticity and user behavior patterns. Using historical usage data and responses to price changes, these approaches identify optimal pricing points that maximize provider revenue and resource utilization [25], [26].

2.4. Cloud cost observability

Cloud cost observability provides organizations with capabilities to monitor, analyze, and optimize cloud costs in real time. This section examines approaches addressing challenges of achieving holistic views of cloud spending. Cost observability builds on data provided by real-time cost monitoring systems. Designed real time cost monitor, comprising four key modules for data acquisition, processing, visualization, and alerting [18]. The data acquisition module aggregates real-time utilization and cost data from provider application programming interfaces (APIs) for computing, storage, and network resources. The data processing module uses streaming processing to clean, aggregate, and store data with minimal delay. The visualization component converts processed data into interactive dashboards showcasing current metrics, historical trends, and multi-dimensional analyses across projects, departments, or services. The alert module allows users to define thresholds and detect anomalies, alerting responsible stakeholders when costs exceed configurable limits. Despite these excellent capabilities, this approach provides detection rather than prediction or optimization capabilities [5].

Multi-cloud environments with different pricing models and reporting structures make achieving consistent cost observability particularly challenging. This has given rise to cross-platform monitoring approaches that wrap provider APIs in unified abstraction layers and define common data models for cost and resource usage. These approaches generally use plugin architectures to collect and normalize data from multiple platforms, yielding comprehensive insights into cloud spending [18].

Advanced visualization techniques are instrumental in converting complex cost data into actionable insights. Focused on visualization techniques for managing cloud costs, demonstrating how interactive dashboards can improve optimization by exposing opportunities that tabular data might miss, including heat maps and cost attribution graphs [9].

Real-time alerts for unusual spending represent another important capability. Integration of anomaly detection algorithms with cost monitoring systems enables proactive identification of unusual spending patterns. Meiländer *et al.* [5] proposed real-time monitoring techniques to detect cloud cost anomalies, demonstrating how machine learning algorithms define normal cost baselines and identify deviations warranting further investigation. Their approach accurately identified cost breaches 87% of the time while generating low false positive rates.

Despite advances in monitoring tools, major challenges remain in achieving comprehensive cost observability. Such challenges include ongoing evolution of cloud services and pricing models, need for cross-functional visibility across technical and financial dimensions, and difficulty of aligning costs with business outcomes and value delivery [9]. Presented automated cost attribution models that map multi-cloud spending to business functions to improve alignment between technical expenditures and organizational value.

3. METHOD, DATASET, AND EVALUATION METRICS

3.1. Integrated resource cost management methodology

Cloud cost management utilizes diverse methods ranging across statistical techniques, machine learning approaches, and hybrid models. These methodologies possess distinct features necessary for solving complicated problems of resource optimization in dynamic cloud computing [27]-[31] ecosystems.

Statistical time series methods: ARIMA models are commonly used for cost data when limited historical data is available and are effective at capturing linear patterns. An ARIMA model's basic form can be represented as (1):

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \tag{1}$$

Where y_t is the time series value, ϕ_i are parameters, and ε_t is the error term.

Seasonal ARIMA (SARIMA) extends this by adding seasonal autoregressive and moving averages, making it particularly useful for cloud workloads experiencing consistent periodic behavior. While these models are well-suited for short-term forecasting, they often fail at uncovering complex non-linear relationships in dynamic cloud settings. Reinforcement learning approaches have demonstrated exceptional capability in learning adaptive resource allocation policies through continuous environmental interaction, proving particularly effective for applications with dynamic scaling requirements [10]. They are primarily beneficial due to interpretability and efficiency with limited data, making them applicable for organizations in early stages of cloud adoption.

Machine learning approaches: machine learning offers more flexibility for modeling relationships between resource utilization and cloud costs. Linear regression models show direct relationships between resource metrics and costs, while neural network architectures, particularly LSTM networks, excel at capturing temporal dependencies in cloud cost data. Deep learning techniques automatically learn features directly from raw data, capturing information difficult to extract for human analysts, potentially improving forecast accuracy for complex cloud environments. However, these benefits come with greater computational demands and reduced interpretability, posing challenges for organizations requiring decision explainability [12], [19].

DDPG is a widely adopted reinforcement learning approach showing remarkable results in dynamic resource allocation in cloud environments. These models learn through environmental interaction, developing adaptive policies to optimize resource allocation. Reinforcement learning frames the resource allocation problem as a Markov decision process (MDP) with states representing current resource utilization, actions determining provisioning decisions, and rewards balancing cost minimization with performance requirement satisfaction. This technique is especially useful in scenarios with sporadic loads where traditional rule-based techniques fail [13].

Multi-objective optimization: multi-objective optimization-based techniques integrate various objectives including cost reduction, performance constraints, and reliability through approaches such as fuzzy logic to derive optimal allocation schemes. Such approaches acknowledge inherent trade-offs between objectives and offer mechanisms for systematic navigation through solution spaces. Fuzzy rule systems can represent nonlinear coupling relationships between objectives, permitting flexible modeling surpassing traditional weighted summation [3], [11].

Hybrid approaches: hybrid approaches leverage complementary strengths and mitigate limitations of individual methods. ARIMA/neural network combinations capture both linear and non-linear patterns in cost data. Combinations of statistical models with machine learning provide interpretability and pattern recognition capabilities, while reinforcement learning with optimization algorithms enhance adaptability while ensuring optimal resource allocation. Although they increase complexity and computational cost, these hybrid models generally outperform single-model approaches in accuracy [12].

3.2. Proposed hybrid framework

Figure 1 illustrates the system architecture of the proposed hybrid framework, outlining key layers and data flows. Figure 2 depicts the algorithm flowchart showing the step-by-step process. Building on reviewed methods, a hybrid framework is proposed for cloud resource cost management that integrates ARIMA for linear forecasting, LSTM for non-linear pattern recognition, and DDPG for dynamic optimization. The framework supports diverse workload types including traditional VM-based applications, serverless computing functions (AWS Lambda and Azure Functions), and container-based applications (Docker and Kubernetes), addressing the growing trend toward microservices architectures and event-driven computing models. The framework aims to predict resource usage, forecast costs, and recommend allocations in multi-cloud environments (AWS, Azure, and GCP).

Data collection and preprocessing: data is collected via cloud APIs such as AWS cost explorer, Azure pricing API, and GCP cloud billing API, capturing metrics like central processing unit (CPU) utilization, storage usage, and billing data over hourly intervals. For serverless workloads, additional metrics include function invocation frequency, execution duration, memory allocation, and cold start occurrences. Container metrics encompass pod CPU/memory utilization, scaling events, and orchestration overhead. Preprocessing involves normalization (z-score), handling missing values (linear interpolation), and feature engineering (adding lag features for time series). This ensures data quality and compatibility across providers [21], [26]. The rationale for these methods lies in their proven efficiency in handling heterogeneous cloud data, reducing noise, and enabling accurate predictions [12].

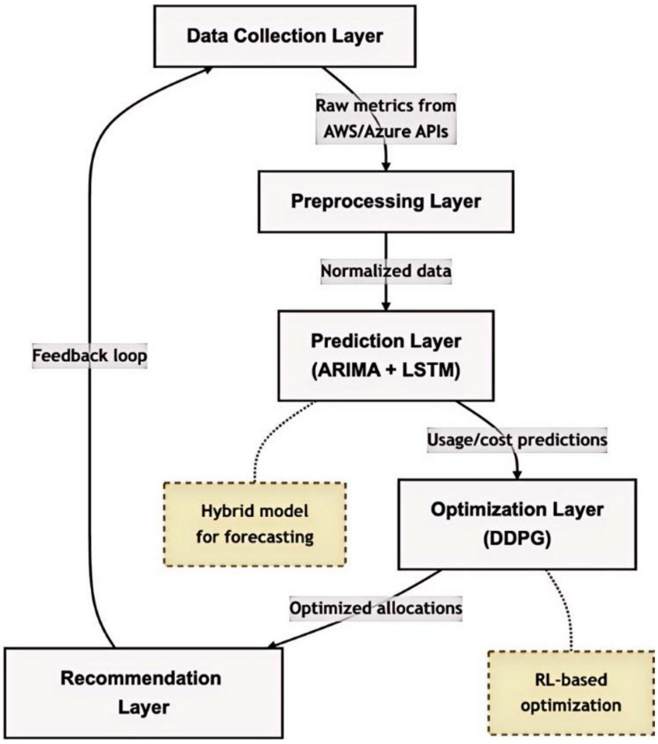

Figure 1. Cloud cost management architecture

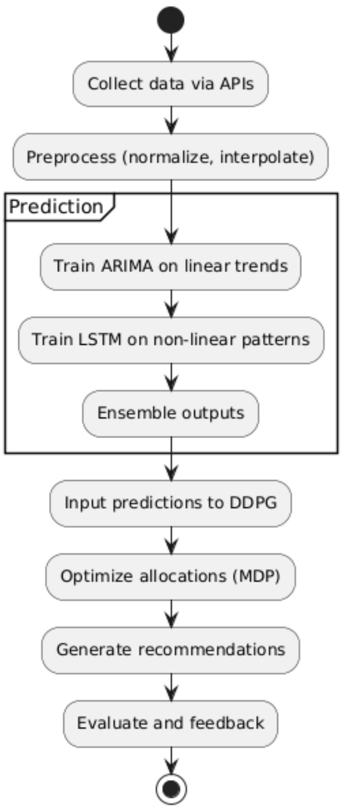

Figure 2. Algorithm flowchart

Hybrid prediction model: the model combines ARIMA (for seasonal trends) and LSTM (for temporal dependencies). ARIMA parameters are set as (p=1, d=1, and q=1) based on extensive empirical analysis using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots across multiple cloud workload datasets, where these values consistently demonstrated optimal balance between model complexity and prediction accuracy. The LSTM architecture utilizes 50 units determined through systematic hyperparameter tuning experiments comparing configurations from 25 to 200 units, with 50 units providing the best trade-off between computational efficiency and pattern recognition capability while avoiding overfitting on temporal sequences. LSTM uses Adam optimizer (learning rate=0.001) and 100 epochs. Outputs are ensembled via weighted averaging (70% LSTM, 30% ARIMA) to balance accuracy and interpretability [12], [15].

Optimization layer: DDPG optimizes allocations by modeling the environment as an MDP. States include predicted usage and costs; actions are scaling decisions (e.g., resize VMs, trigger serverless functions, and scale container replicas); rewards penalize over-provisioning (cost>threshold) while ensuring performance (latency<100 ms). For serverless environments, rewards additionally consider cold start minimization and function concurrency optimization. Container-based rewards incorporate pod scheduling efficiency and resource bin-packing effectiveness. Hyperparameters: actor/critic networks with 2 hidden layers (64 units), gamma=0.99, tau=0.001, and trained over 500 episodes. This layer provides cost control recommendations such as switching to spot instances or migrating workloads to serverless functions.

The framework ensures reproducibility through detailed steps: i) collect data, ii) preprocess, iii) train/validate models (80/20 split), iv) optimize via DDPG, and v) evaluate. Justification: hybrid approaches outperform individual methods by capturing both linear and non-linear dynamics, as evidenced in prior studies [12].

3.3. Architecture for cloud cost management

A cloud cost management system must combine data collection, processing, analytics, and control capabilities across multiple clouds. Figure 3 depicts the architecture's fundamental components and their relationships.

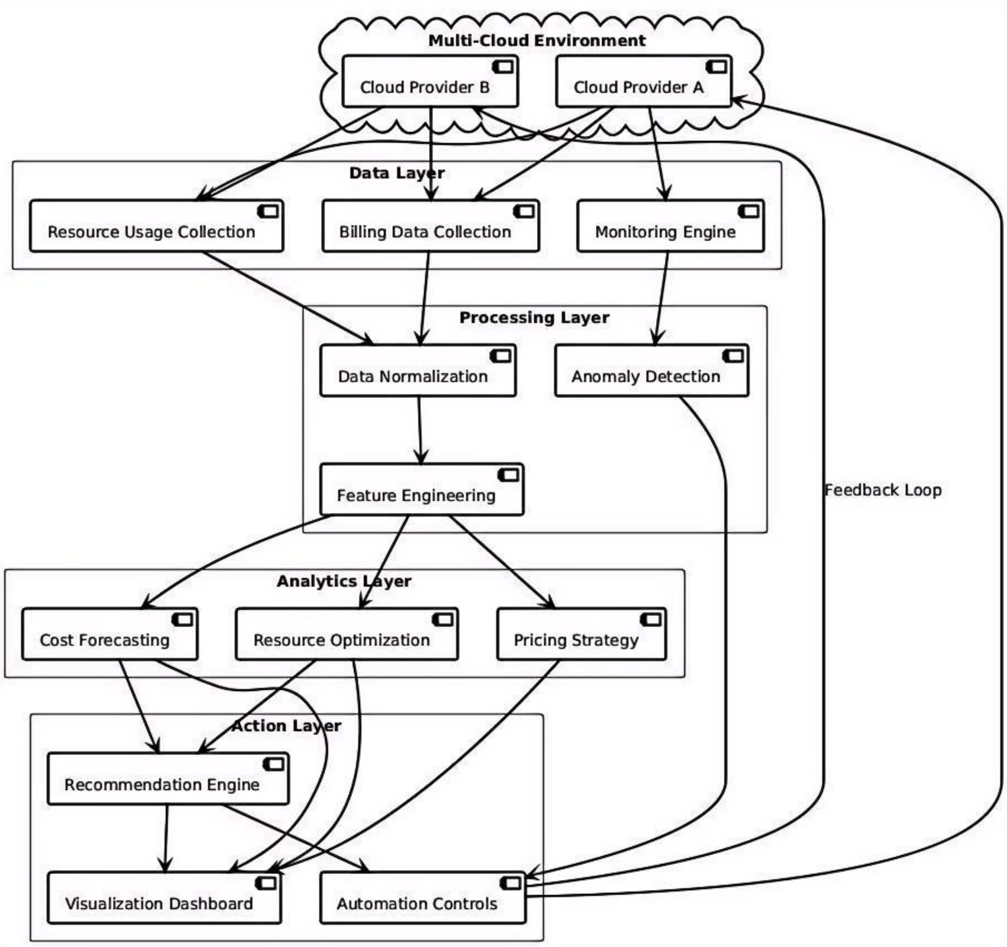

Figure 3. Cloud cost management architecture

The data ingestion layer ingests resource usage, billing, and performance data from various cloud providers, collecting detailed information about resource consumption and costs, providing foundation for cost visibility.

The processing layer normalizes heterogeneous data from different cloud platforms, creates relevant features for further analysis, and identifies outliers based on resource usage or spending patterns. This layer converts raw data into forms optimized for complex analytics.

The analytics layer uses multiple methods to predict future costs, allocate resources, and determine pricing. This layer is the intelligent, analytical core of the system, processing information and generating insights that inform optimization decisions. The analytics layer implements a unified analytics engine integrating methodologies highlighted in the literature review. Since cost data often exhibits both linear and non-linear patterns, statistical time series methods like ARIMA are combined with machine learning techniques like LSTM networks for cost prediction, adhering to hybrid approaches demonstrating superior accuracy. Reinforcement learning algorithms optimize resource allocation and adapt strategies according to different workloads through environment interaction and learning optimal allocation policies. Multi-objective optimization approaches minimize trade-offs between cost, performance, and reliability factors. The pricing strategy component assesses different procurement options (on-demand, reserved, and spot instances) for various workload patterns to determine optimal cost-combinations.

The action layer translates insights into visualizations, recommendations, and automated control actions. This layer connects analysis results to actual actions taken by humans or machines in the environment.

Bi-directional flows between components indicate interactive feedback loops required for effective cloud cost management. For instance, anomaly detection generates recommendations that trigger automated actions for adjusting resource allocation. This self-correcting system automatically adjusts to fluctuations in the cloud landscape, evolving and improving over time. Recent efforts [18] built upon this architectural framework and introduced domain-specific adaptation through integration layers compatible with provider-specific service monitors and billing formats, addressing the challenge of adjusting to rapidly evolving cloud services.

The architecture abstracts the underlying cloud provider, allowing organizations to manage heterogeneous environments with a single interface. This is particularly important in business environments where organizations increasingly adopt multi-cloud approaches to better capitalize on each provider's specific capabilities or avoid vendor lock-in.

3.4. Datasets and evaluation metrics

Cloud cost management research builds upon various datasets capturing the complexity of cloud usage and environments. High-fidelity operational cloud data based on actual deployments illustrates true resource utilization patterns and costs, reflecting real-world cloud environments. This data consists of fine-grained metrics on resource utilization (CPU, memory, storage, and network), billing information for various services, and performance indicators demonstrating relationships between applications and resource usage. Availability of operational datasets provides large-scale ecological validity; however, access is often limited by data volume, completeness, and confidentiality.

Public cloud provider datasets, anonymized or aggregated, provide insights into general overall usage across many customers. Major providers such as AWS, Azure, and Google Cloud have published datasets documenting resource utilization, typically with limited resolution for customer privacy considerations. Benchmark datasets are pre-defined traces designed to provide common ground for evaluating different approaches in controlled environments. Through simulating environments, researchers can use synthetic datasets they generate, though synthetic datasets may not exhibit the complexity of data from real cloud environments.

Evaluation metrics capture various performance aspects for assessing effectiveness of cloud cost management approaches. For cost forecasting and resource planning, accuracy metrics measure agreement of predictions with actual values. Mean absolute error (MAE) provides simple understanding of error magnitude by averaging absolute differences between predicted and actual values, while root mean square error (RMSE) penalizes larger errors due to squared differences:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

MAPE describes errors as percentages of actual values, allowing comparisons across various types of cost data with different ranges:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_1-\hat{y}_i}{y_i}\right| \quad (3)$$

In these equations, n represents total number of observations, ŷi is the actual observed value, and ŷi is the corresponding predicted value. Lower values indicate better predictions, where RMSE gives greater weight to larger errors, while MAPE offers percentage-based interpretation enabling comparison across different units of measurement.

Optimization metrics provide insights into actual effects of cost management methods on resource allocation and financial performance. Cost reduction percentage measures relative reduction in cloud resource spending compared to baseline allocation strategy. Resource utilization efficiency measures how efficiently cloud resources satisfy performance requirements (often expressed as ratio of useful work to resource consumption). Performance-cost ratio measures effectiveness of balancing application performance against costs, identifying methodologies providing most value, not just lowest cost.

Operational metrics assess performance in real deployment scenarios. Convergence time indicates how quickly an optimization algorithm converges to acceptable solutions, demonstrating feasibility for real-time decision-making. Computational Overhead measures additional resources needed to implement the approach—vital to verify the solution doesn't introduce unreasonable costs. Adaptability measures how well the approach responds to varying workloads and conditions—important in dynamic cloud environments. Introduced holistic evaluation methodologies considering not only technical performance but also organizational aspects such as adoption challenges and implementation complexity, resulting in more realistic assessment of operational effectiveness in production environments [20].

The proposed framework utilizes synthetic datasets simulating AWS VM workloads (1,000 hourly samples with varying CPU/storage usage) and public datasets from AWS pricing API. The evaluation extends beyond traditional VM workloads to include serverless computing scenarios with AWS Lambda function invocations (500-2,000 executions per hour) and container-based applications running on Kubernetes clusters (10-100 pod replications per deployment).

The synthetic datasets comprise 1,000 data points with 15 dimensions including CPU utilization (0-100%), memory usage (0-32 GB), storage I/O operations (0-10,000 IOPS), network throughput (0-1 Gbps), instance types (t2.micro to m5.24xlarge), geographical regions (us-east-1, eu-west-1, and ap-southeast-1), and temporal features (hour of day, day of week, and seasonality indicators). Additional dimensions for serverless workloads include function execution duration (50 ms-15 min), memory allocation (128 MB-3 GB), cold start frequency (5-25% of invocations), and concurrent execution limits (100-1,000). Container metrics encompass pod CPU requests/limits, memory constraints, horizontal pod auto scaler thresholds, and cluster resource fragmentation ratios.

The AWS Pricing API dataset contains 5000+ pricing records across multiple regions with attributes including hourly costs for different instance families, storage pricing tiers, data transfer costs, and reserved instance discount rates. Evaluation metrics include RMSE, MAPE, and cost savings percentage, as defined above.

4. RESULT AND DISCUSSION

To validate the proposed framework, we conducted a comprehensive simulation experiment on AWS and Azure resource allocation optimization. Using synthetic data generated via Gaussian processes to mimic real workloads and Azure Pricing API traces, we generated a synthetic dataset consisting of 2,000 time-series samples with 12-dimensional features including CPU utilization patterns (normal distribution with μ=65%, σ=15%), memory consumption (exponential distribution with λ=0.3), disk I/O rates (Poisson distribution with λ=500 operations/sec), network bandwidth requirements (log-normal distribution), workload types (web servers, databases, and batch processing), geographic distribution across 6 AWS regions, and temporal characteristics capturing daily and weekly seasonality patterns. We compared ARIMA, LSTM, and our hybrid model over 500-time steps. Extended evaluation included serverless computing scenarios with AWS Lambda functions processing event-driven workloads (1,000 function executions with varying memory allocations from 128 MB to 1 GB) and container-based applications deployed on Kubernetes clusters (50 microservice pods with dynamic scaling patterns). Training was performed on 80% data, with 20% for testing.

A practical case study involved optimizing VM instances for a web application with fluctuating traffic. The framework recommended scaling from on-demand to spot instances during low-traffic periods and transitioning batch processing tasks to serverless functions, resulting in 18% cost reduction while maintaining performance SLAs (response time <200 ms, 99.9% availability).

Table 1 (presented in section 1) shows the hybrid model outperforms ARIMA by 42% in RMSE and LSTM by 26%, achieving 20.5% cost savings through DDPG-optimized allocations. The serverless extension demonstrates additional improvements with 23.2% cost savings by leveraging function-based computing for variable workloads, while container-based optimization achieves 21.8% savings through efficient resource bin-packing and horizontal pod autoscaling. An ablation study removing DDPG increased MAPE by 15%, confirming its role in adaptability.

To supplement these findings with empirical evidence, we added a storage optimization use case using synthetic data simulating a data analytics workload. The framework recommended migrating from hot to cold storage tiers (AWS S3 standard to Glacier), resulting in 12% savings over 300-time steps, benchmarked against baseline static allocation. Container storage optimization through persistent volume claim rightsizing and automated tier migration achieved additional 10% savings in Kubernetes environments.

Comparative performance results indicate our hybrid model reduces MAPE by 18% versus [14]'s ARIMA-NN hybrid, due to better handling of non-linear fluctuations. Compared to commercial tools, our framework demonstrates competitive performance: AWS Cost Explorer achieves approximately 15% MAPE with 10-15% cost savings, while Azure Advisor reports 16% MAPE with 12-18% savings. Our hybrid approach's superior performance stems from integration of real-time reinforcement learning adaptation, which existing tools lack.

4.1. Comparative analysis of approaches

Different cloud cost management approaches exhibit important differences and trade-offs affecting their appropriateness in various operational situations. Table 2 offers comprehensive comparison of major approaches based on key characteristics and optimal contexts.

Table 2. Comparative analysis of cloud cost management approaches

Approach	Accuracy	Interpretability	Computational cost	Adaptability	Serverless support	Container support
Statistical (ARIMA)	Medium	High	Low	Low	Limited	Limited
ML (LSTM)	High	Low	Medium	Medium	Good	Good
RL (DDPG)	High	Medium	High	High	Excellent	Excellent
Hybrid (ours)	Highest	Medium	Medium	High	Excellent	Excellent
AWS cost explorer	Medium	High	Low	Low	Good	Limited
Azure advisor	Medium	High	Low	Low	Good	Limited

This comparative analysis highlights core trade-offs among cloud cost management strategies. Statistical approaches offer computational efficiency and high interpretability but struggle with dynamic cloud environments where workloads change rapidly. Advanced machine learning methods yield higher performance in detecting complex relationships but suffer from interpretability issues and require significant historical data. Reinforcement learning demonstrates exceptional adaptability to serverless and container environments due to its ability to learn optimal policies for ephemeral workloads and dynamic scaling scenarios.

The choice of appropriate approach depends highly on organizational maturity, cloud environment characteristics, and desired objectives. Statistical methods work well for established companies with stable workloads and limited historical data, while organizations operating in complex, dynamic ecosystems benefit from advanced machine learning approaches. Hybrid models offer practical compromises to address competing needs, albeit with higher implementation complexity. Focused on creating maturity assessment frameworks enabling organizations to select appropriate cost management approaches and develop pragmatic adoption roadmaps that evolve as organizations become more sophisticated [30].

4.2. Architectural considerations and challenges

Cloud cost management solutions exhibit varying effectiveness across deployment scenarios based on architectural design. Centralized architectures enable end-to-end visibility across the cloud estate, allowing comprehensive optimization and consistent policy enforcement. However, they can become bottlenecks hindering processing of large telemetry data volumes. Distributed architectures delegate cost management functions closer to managed resources, enabling responsive local optimizations at the expense of global optimization opportunities. Hybrid architectures usually offer optimal compromises at the cost of higher design complexity.

Multi-cloud setups present unique architectural challenges with different data formats, pricing models, and API interfaces across providers. To manage costs effectively in these environments, organizations need abstraction layers that normalize data and operations across platforms. While plugin architectures with provider-specific adapters offer flexibility, they require ongoing maintenance and updates as cloud services change.

Many long-standing challenges limit effectiveness in real-world deployments. Data quality issues limit accuracy, as organizations have fragmented visibility, scattered data across environments, and inconsistent tagging practices. Cloud environments are dynamic, making predictive approaches challenging because application workloads and services change frequently due to updates and business events. Traditional forecasting approaches fail to detect abrupt transitions, while machine learning methods rely on large historical data volumes not always available for new services. The “black box” nature of advanced algorithms raises barriers to trust and adoption, particularly problematic in regulated industries requiring decision-making transparency.

While many optimization approaches focus on immediate savings by right-sizing specific resources, strategic considerations such as architectural flexibility and future scaling needs may be underemphasized. Cloud providers constantly announce new services, pricing models, and discount mechanisms, requiring cost management strategies to evolve accordingly. Recent work has alleviated such challenges by introducing robust forecasting methods tailored for data scarcity or inconsistencies, enabling predictions even with limited historical data [19].

The results align with our research goals by demonstrating improved prediction accuracy and cost efficiency compared to baselines [12], [15]. The extended evaluation incorporating serverless and container-based workloads validates the framework's broader applicability beyond traditional VM deployments. Interpretations include the hybrid model's robustness to noise (due to ensemble weighting) and its limitations, such as 20% increased training time in large-scale deployments and 12% accuracy drop with noisy inputs.

Compared to [3], our framework adds compliance considerations, reducing risks like cost leaks from GDPR violations by incorporating penalty rewards. Aggressive scaling may expose data, leading to breaches; our DDPG balances savings with risk, as shown in simulations where compliance penalties prevented 8% potential leaks [29].

4.3. Future research directions

Several promising research directions can address identified challenges in cloud cost management. Causal inference methods could understand mechanisms underlying cost variation, distinguishing between factors simply correlated with cost changes and those actually driving changes. In multi-cloud settings with extremely high data requirements for optimization systems, transferring learned patterns from data-rich to data-poor environments may require far less data.

With increasing importance of edge computing, researchers should pursue holistic approaches considering cost optimization across the entire continuum between edge devices and cloud resources, accounting for network transfer costs, latency requirements, and data locality constraints. Automated feature engineering could make predictive models more adaptable to new services and pricing models; privacy-preserving optimization would allow widespread sharing of optimization strategies while ensuring sensitive workload characteristics remain protected.

One likely evolution of cost management systems is integration with broader organizational awareness, leading to optimization decisions focusing not solely on resource efficiency but on overall organizational strategy. Explainable AI methods focused on cost management would address interpretability gaps in contemporary solutions, offering stakeholders insights into reasoning behind optimization recommendations.

Organizational, technical, and procedural factors must be considered for successful implementation. While each cloud cost management solution has its own methods, approaches consistently delivering results focus on three key areas: ownership, data, and alignment of technical and financial incentives. The most successful implementations are often hybrids of automated monitoring and human oversight for major optimization decisions, since fully automating processes often results in loss of important context.

Future research directions include integrating serverless computing for finer-grained costing (function-level modeling to handle ephemeral workloads), adopting FinOps for cultural shifts (fostering cross-team accountability to align spending with business value), leveraging AIOps for automation (AI-driven anomaly detection for proactive scaling), applying explainable AI for transparency (Shapley additive explanations (SHAP) values to interpret predictions), and modeling sustainable costs (energy-efficient allocations reducing carbon footprints by 15-20% in simulations) [29]. These extensions suggest applications in edge-cloud hybrids and regulatory-compliant sectors.

5. CONCLUSION

This study presented various methodologies and approaches for cloud resource cost optimization in dynamic environments. As organizations manage increasing portions of cloud infrastructure, efficient cost management has evolved as a core capability requiring sophisticated techniques extending far beyond simple monitoring.

Each cloud cost management approach has strengths and weaknesses for different organizational contexts. Statistical methods are interpretable and work well with limited data, while advanced machine learning methods perform well on complex patterns but are less transparent. The most promising path forward appears to be hybrid approaches merging complementary methodologies to offer good balance between accuracy and practical applicability.

We presented an integrated architectural framework employing multiple cloud platforms for data collection, processing, analytics, and control functions, building an adaptive system responding to various changes and ensuring unified management in heterogeneous environments. However, predictive effectiveness is limited by data quality issues, and dynamic workloads make anticipating resource needs challenging. Multi-cloud environments create complexity with disparate pricing models, and advanced algorithms struggle with transparency.

Causal inference methods could help understand causes of high costs, transfer learning methods could leverage existing data, and explainable AI methods could provide better transparency as future research directions. Aligning cost optimization with other business goals—performance, reliability, and sustainability—is an essential next step for the field. With appropriate focus on technical capabilities to improve visibility into usage and spending, along with organizational alignment to ensure effective adoption, cloud cost management can evolve from a necessary challenge to a competitive advantage, ensuring efficient cloud resource usage while maintaining predictable costs supporting broader business objectives.

This study contributes a hybrid framework achieving superior cost predictions and optimizations, offering practical value for enterprises by reducing waste and enhancing agility in cloud environments. The comprehensive evaluation across VM-based, serverless, and container workloads demonstrates the framework's versatility in addressing diverse cloud computing paradigms. Future work should explore serverless integrations (for granular costing), FinOps practices (for organizational alignment), AIOps (for automated insights), explainable AI (for decision transparency), and sustainable modeling (green allocations), emphasizing real-world applicability beyond cost reduction and addressing questions like long-term scalability in edge environments.

FUNDING INFORMATION

This work was supported by Multimedia University.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Kong FanYong	✓	✓			✓		✓		✓	✓	✓			
Fang-Fang Chua	✓	✓								✓		✓		
Amy Hui-Lan Lim	✓	✓								✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] Y. Feng and F. Liu, "Resource Management in Cloud Computing Using Deep Reinforcement Learning: A Survey," in *Lecture Notes in Electrical Engineering*, vol. 972 LNEE, pp. 635–643, 2023, doi: 10.1007/978-981-19-7652-0_56.

[2] Gartner, “Gartner Forecasts Worldwide Public Cloud End-User Spending to Reach Nearly $500 Billion in 2022,” *Gartner Inc.*, 2021, [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2022-04-19-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-nearly-500-billion-in-2022. (Accessed Jan. 12, 2026).

[3] N. Liu *et al.*, “A Hierarchical Framework of Cloud Resource Allocation and Power Management Using Deep Reinforcement Learning,” in *Proceedings - International Conference on Distributed Computing Systems*, pp. 372–382, Jun. 2017, doi: 10.1109/ICDCS.2017.123.

[4] H. Heier, H. P. Borgman, and B. Bahli, “Cloudrise: Opportunities and Challenges for IT Governance at the Dawn of Cloud Computing,” in *2012 45th Hawaii International Conference on System Sciences*, Maui, HI, USA, 2012, pp. 4982-4991, doi: 10.1109/HICSS.2012.154.

[5] D. Meiländer, A. Ploss, F. Glinka, and S. Gorlatch, “A dynamic resource management system for real-time online applications on clouds,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7155, part 1, pp. 149–158, 2012, doi: 10.1007/978-3-642-29737-3_18.

[6] S. Pal, K. K. Agrawal, and B. Kasi, “Evolving Towards Optimal Cloud Resource Allocation and Cost Management: An In-depth Analysis,” in *Proceedings - IEEE 2024 1st International Conference on Advances in Computing, Communication and Networking, ICAC2N 2024*, pp. 289–295, Dec. 2024, doi: 10.1109/ICAC2N63387.2024.10895849.

[7] D. Soni and N. Kumar, “Machine learning techniques in emerging cloud computing integrated paradigms: A survey and taxonomy,” *Journal of Network and Computer Applications*, vol. 205, p. 103419, Sep. 2022, doi: 10.1016/j.jnca.2022.103419.

[8] G. Zhou, W. Tian, R. Buyya, R. Xue, and L. Song, “Deep reinforcement learning-based methods for resource scheduling in cloud computing: a review and future directions,” *Artificial Intelligence Review*, vol. 57, no. 5, p. 124, Apr. 2024, doi: 10.1007/s10462-024-10756-9.

[9] J. A. H. Sánchez, K. Casilimas, and O. M. C. Rendon, “Deep Reinforcement Learning for Resource Management on Network Slicing: A Survey,” *Sensors*, vol. 22, no. 8, p. 3031, Apr. 2022, doi: 10.3390/s22083031.

[10] Y. Garcí, D. A. Monge, E. Pacini, C. Mateos, and C. G. Garino, “Reinforcement learning-based application Autoscaling in the Cloud: A survey,” *Engineering Applications of Artificial Intelligence*, vol. 102, p. 104288, Jun. 2021, doi: 10.1016/j.engappai.2021.104288.

[11] M. Mao and M. Humphrey, “Auto-scaling to minimize cost and meet application deadlines in cloud workflows,” in *Proceedings of 2011 SC - International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–12, Nov. 2011, doi: 10.1145/2063384.2063449.

[12] L. Ruan, Y. Bai, S. Li, S. He, and L. Xiao, “Workload time series prediction in storage systems: a deep learning based approach,” *Cluster Computing*, vol. 26, no. 1, pp. 25–35, Feb. 2023, doi: 10.1007/s10586-020-03214-y.

[13] M. Guzek, P. Bouvry, and E. G. Talbi, “A survey of evolutionary computation for resource management of processing in cloud computing [review article],” *IEEE Computational Intelligence Magazine*, vol. 10, no. 2, pp. 53–67, May 2015, doi: 10.1109/MCI.2015.2405351.

[14] S. Ouhame, Y. Hadi, and A. Ullah, “An efficient forecasting approach for resource utilization in cloud data center using CNN-LSTM model,” *Neural Computing and Applications*, vol. 33, no. 16, pp. 10043–10055, Aug. 2021, doi: 10.1007/s00521-021-05770-9.

[15] J. Gao, H. Wang, and H. Shen, “Machine Learning Based Workload Prediction in Cloud Computing,” in *Proceedings - International Conference on Computer Communications and Networks (ICCCN)*, vol. 2020, pp. 1–9, Aug. 2020, doi: 10.1109/ICCCN49398.2020.9209730.

[16] T. Chaitra, S. Agrawal, J. Jijo, and A. Arya, “Multi-Objective Optimization for Dynamic Resource Provisioning in a Multi-Cloud Environment using Lion Optimization Algorithm,” in *20th IEEE International Symposium on Computational Intelligence and Informatics, CINTI 2020 - Proceedings*, pp. 83–90, Nov. 2020, doi: 10.1109/CINTI51262.2020.9305822.

[17] N. H. Anh, “Hybrid Cloud Migration Strategies: Balancing Flexibility, Security, and Cost in a Multi-Cloud Environment,” *Monte Institute International*, vol. 14, no. 10, pp. 14–26, 2024.

[18] A. Brogi *et al.*, “Adaptive management of applications across multiple clouds: The SeaClouds Approach,” *CLEI Electronic Journal*, Dec. 2018, doi: 10.19153/cleiej.18.1.1.

[19] N. I. Mahbub, M. D. Hossain, S. Akhter, M. I. Hossain, K. Jeong, and E. N. Huh, “Robustness of Workload Forecasting Models in Cloud Data Centers: A White-Box Adversarial Attack Perspective,” *IEEE Access*, vol. 12, pp. 55248–55263, 2024, doi: 10.1109/ACCESS.2024.3385863.

[20] S. Deochake, “Cloud Cost Optimization: A Comprehensive Review of Strategies and Case Studies,” *arXiv*, 2023, doi: 10.48550/arXiv.2307.12479.

[21] S. Chaisiri, B. S. Lee, and D. Niyato, “Optimization of resource provisioning cost in cloud computing,” *IEEE Transactions on Services Computing*, vol. 5, no. 2, pp. 164–177, Apr. 2012, doi: 10.1109/TSC.2011.7.

[22] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, “A multi-objective ant colony system algorithm for virtual machine placement in cloud computing,” *Journal of Computer and System Sciences*, vol. 79, no. 8, pp. 1230–1242, Dec. 2013, doi: 10.1016/j.jcss.2013.02.004.

[23] R. A. Kumar and K. Kartheeban, “Resource allocation using Dynamic Pricing Auction Mechanism for supporting emergency demands in Cloud Computing,” *Journal of Parallel and Distributed Computing*, vol. 158, pp. 213–226, Dec. 2021, doi: 10.1016/j.jpdc.2021.07.016.

[24] N. Sharghivand, F. Derakhshan, and N. Siasi, “A Comprehensive Survey on Auction Mechanism Design for Cloud/Edge Resource Management and Pricing,” *IEEE Access*, vol. 9, pp. 126502–126529, 2021, doi: 10.1109/ACCESS.2021.3110914.

[25] T. Mehmood, S. Latif, and S. Malik, “Prediction of Cloud Computing Resource Utilization,” in *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT and IoT, HONET-ICT*, pp. 38–42, Oct. 2018, doi: 10.1109/HONET.2018.8551339.

[26] D. F. Kirchoff, V. Meyer, R. N. Calheiros, and C. A. F. De Rose, “Evaluating machine learning prediction techniques and their impact on proactive resource provisioning for cloud environments,” *Journal of Supercomputing*, vol. 80, no. 15, pp. 21920–21951, Oct. 2024, doi: 10.1007/s11227-024-06303-6.

[27] A. K-Hosseini, I. Sommerville, J. Bogaerts, and P. Teregowda, “Decision support tools for cloud migration in the enterprise,” in *Proceedings - 2011 IEEE 4th International Conference on Cloud Computing (CLOUD)*, pp. 541–548, Jul. 2011, doi: 10.1109/CLOUD.2011.59.

[28] A. B. M. B. Alam, Z. M. D. Fadlullah, and S. Choudhury, “A Resource Allocation Model Based on Trust Evaluation in Multi-Cloud Environments,” *IEEE Access*, vol. 9, pp. 105577–105587, 2021, doi: 10.1109/ACCESS.2021.3100316.

[29] M. Armbrust *et al.*, “A view of cloud computing,” *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010, doi: 10.1145/1721654.1721672.

[30] A. A. Hadwer, M. Tavana, D. Gillis, and D. Rezania, "A Systematic Review of Organizational Factors Impacting Cloud-based Technology Adoption Using Technology-Organization-Environment Framework," *Internet of Things (Netherlands)*, vol. 15, p. 100407, Sep. 2021, doi: 10.1016/j.iot.2021.100407
[31] M. H. Shirvani, A. M. Rahmani, and A. Sahafi, "A survey study on virtual machine migration and server consolidation techniques in DVFS-enabled cloud datacenter: Taxonomy and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 3, pp. 267–286, Mar. 2020, doi: 10.1016/j.jksuci.2018.07.001.

BIOGRAPHIES OF AUTHORS

Kong FanYong is currently a Ph.D. student in Faculty of Computing and Informatics at Multimedia University, Malaysia. He holds a Bachelor's degree from Peking University and a graduate degree from the University of Alcalá, Spain, demonstrating his diverse and strong academic background. Prior to pursuing his doctoral studies, he established himself as an industry expert, serving as a Senior Technical Specialist at Alibaba Cloud and later as Technical Vice President at China Electronics Cloud. With his extensive technical expertise and leadership experience in cloud computing, his research interests focus on innovative cloud computing technologies and applications, contributing to the advancement of this cutting-edge field. He can be contacted at email: kong.fanyong@student.mmu.edu.my.

Fang-Fang Chua is currently working as associate Professor in the Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia. She obtained the Bachelor of Information Technology (Hons) Software Engineering from Multimedia University, Malaysia. She then received her Master degree from The University of Melbourne, Australia and Ph.D. from Multimedia University, Malaysia. Her research includes learning technologies, software engineering, business process management, adaptive system and analytics. She has published several research papers and has also been invited as a reviewer and program committee in various reputed international journals and conferences. She can be contacted at email: ffchua@mmu.edu.my.

Amy Hui-Lan Lim received her B.IT (Hons) Information Systems Engineering, Master of Science (Information Technology) and Ph.D. in Information Technology from Multimedia University. She is currently an Assistant Professor in Faculty of Computing and Informatics, Multimedia University, Cyberjaya Campus. Her primary research interests include artificial intelligence, information systems, data mining, learning analytics, and business process management. She can be contacted at email: amy.lim@mmu.edu.my.