International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Using AI to automatically analyze workload patterns and suggest optimal VM/container sizes, avoiding overprovisioning

Hema Vamsi Nikhil Katakam

Software Development Engineer

Abstract:
In cloud computing, organizations often allocate resources conservatively to guarantee application performance under peak load conditions. However, this practice results in persistent over-provisioning, wasted cost, and increased carbon footprint. This paper proposes an AI-driven resource right-sizing framework that leverages workload telemetry, predictive analytics, and feedback-based optimization to automatically determine optimal configurations for virtual machines (VMs) and containers. Using long short-term memory (LSTM) neural networks, the system forecasts resource demand and dynamically recommends suitable compute, memory, and I/O configurations. The proposed model demonstrates substantial cost savings (30–40%) and improved utilization stability without compromising service-level agreements (SLAs).

Keywords: right-sizing, cloud computing, virtual machine sizing, container sizing, workload prediction, machine learning, resource optimization, cost efficiency, autoscaling.

1. INTRODUCTION

Cloud computing has transformed the way organizations deploy and manage computing resources by providing on-demand access to scalable infrastructure. Platforms such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) allow enterprises to dynamically allocate virtual machines (VMs) and containers as required. However, despite this elasticity, most organizations continue to over-provision resources to ensure performance stability during traffic surges. This practice, though safe, leads to substantial cost inefficiency and wasted computational capacity.

Studies indicate that nearly 30–50 percent of cloud resources remain under-utilized at any given time, resulting in unnecessary spending and higher energy consumption. Manually determining the optimal instance size—known as *right-sizing*—is tedious and error-prone, as administrators must analyze large volumes of telemetry data and anticipate variable workload patterns. The complexity of modern applications further complicates this process: some workloads are constant, others cyclic, and many highly bursty, each demanding different scaling strategies.

Traditional rule-based autoscaling mechanisms, such as AWS Auto Scaling Groups or the Kubernetes Horizontal Pod Autoscaler, react to current utilization using fixed thresholds (e.g., CPU > 70 percent). While these methods provide basic elasticity, they remain reactive rather than predictive. They often trigger late or oscillate frequently, a phenomenon known as “thrashing.” To overcome these limitations, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as enablers of proactive right-sizing, where resource adjustments are guided by data-driven forecasts instead of static rules.

AI-based approaches learn from historical and real-time telemetry—including CPU, memory, and I/O metrics—to forecast upcoming demand and recommend appropriate configurations. Predictive models such as Long Short-Term Memory (LSTM) networks capture temporal dependencies, allowing systems to scale up before an expected spike and scale down during idle periods. When integrated with cloud provider

APIs, these models can automatically suggest or execute resizing actions—for example, reducing a VM from t3.large to t3.medium, or adjusting container resource limits to match forecasted usage.
Beyond cost reduction, intelligent right-sizing advances sustainability. Data centers consume roughly one percent of global electricity, and eliminating idle capacity contributes directly to lower carbon emissions. Thus, optimizing cloud resources supports both economic and environmental goals, aligning with the United Nations' Sustainable Development Goal 12 (Responsible Consumption and Production).
This paper introduces an Intelligent Resource Right-Sizing Architecture (IRSA) that integrates AI-based workload prediction, optimization, and feedback.

2. LITERATURE REVIEW

Efficient resource allocation is central to cloud computing economics. The evolution from static provisioning toward intelligent, data-driven optimization has inspired extensive academic and industrial research. This section reviews the foundational work on (a) right-sizing and cost optimization in cloud environments, (b) workload characterization and autoscaling, and (c) applications of artificial intelligence in resource management.

2.1. Right-Sizing and Cost Optimization in Cloud Environments

Right-sizing refers to the process of matching a virtual machine's (VM) or container's allocated capacity to its actual workload requirements [1] [2]. Early efforts were largely rule-based, relying on periodic human audits or threshold triggers. According to Ahead and Hystax reports, organizations frequently oversubscribe compute resources to guarantee uptime, producing idle servers that inflate operational expenses by 30 – 40 percent. Cloud providers have acknowledged this issue and introduced in-built recommendation tools—such as AWS Compute Optimizer and Azure Advisor—that analyze utilization histories to suggest alternate instance types.

2.2. Workload Pattern Analysis and Autoscaling

Understanding workload behavior is a prerequisite for intelligent resource management. Workloads in cloud data centers can be broadly categorized as steady, periodic, or bursty. Each pattern exhibits distinct temporal characteristics that influence scaling policies.

- Steady workloads (e.g., database or storage services) maintain near-constant utilization; right-sizing mainly involves eliminating historical over-allocation.
- Periodic workloads (e.g., business transactions) follow predictable diurnal or weekly cycles; they benefit from scheduled scaling or predictive allocation.
- Bursty workloads (e.g., streaming analytics or social media) exhibit sudden spikes and require adaptive models with quick feedback loops.

Kubernetes' Horizontal Pod Autoscaler (HPA) exemplifies threshold-based elasticity. It adjusts replica counts based on average CPU or memory usage relative to target values. While simple and widely adopted, HPAs depend on reactive thresholds and lack the foresight to anticipate spikes. Studies highlight that purely reactive systems often oscillate, increasing response latency and causing transient under- or over-utilization [3].
To mitigate these drawbacks, predictive autoscaling was introduced. The regression models forecast workload intensity in cloud clusters, showing up to 25 percent improvement in stability compared with reactive scaling [4]. More recent frameworks, such as Google's proactive workload management, integrate historical trend analysis with autoscaler policies to pre-allocate resources before expected demand surges. Despite progress, such systems remain primarily rule-enhanced rather than fully intelligent; they still require manual parameter tuning and lack generalization across workloads.

2.3. Artificial Intelligence and Machine Learning in Resource Management

Artificial Intelligence (AI) offers a paradigm shift from threshold-driven automation to cognitive decision-making. Machine learning algorithms learn workload signatures from telemetry—CPU, memory, disk I/O, and network metrics—and infer future demand or performance bottlenecks.

Supervised learning techniques, including support vector regression, random forests, and gradient boosting, have been applied for capacity planning. These models achieve reasonable accuracy but require handcrafted features and often fail to capture sequential dependencies.

Time-series deep learning approaches overcome this limitation. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are particularly effective for learning temporal correlations in workload traces [5]. Using LSTM-based predictor for virtual machine scaling the work achieved 93 percent prediction accuracy in synthetic benchmarks. Similarly, another works such as reinforcement-learning framework that jointly optimized energy consumption and response latency, demonstrating dynamic adaptation to unseen workload types.

Reinforcement learning (RL) methods conceptualize resource management as a continuous control problem, where an agent receives a reward for minimizing cost while maintaining SLA compliance [6]. The research, Q-learning-based scheduler that learned optimal scaling actions in fluctuating environments, outperforming static heuristics by 20 percent in cost efficiency [7]. The HUNTER framework combined graph neural networks and deep RL to holistically manage CPU, memory, and power states across large data centers [8]. These AI-driven models highlight the feasibility of autonomic cloud systems capable of self-optimization.

2.4. Feedback Loops and Continuous Optimization

Intelligent right-sizing is not a one-time adjustment but a continuous learning process. Effective systems incorporate a feedback loop that compares predicted versus actual resource utilization and retrains models to minimize forecast error. This aligns with the principle of autonomic computing proposed, where systems self-configure, self-optimize, and self-heal [9].

Feedback-driven architectures have been studied in hybrid environments. For example, a feedback-enhanced ML model that reduced over-provisioning by 37 percent in container clusters while keeping SLA violations under 2 percent. Adaptive loops also enable transfer learning—leveraging patterns learned from one workload to bootstrap another—reducing cold-start overhead for new deployments [10].

Modern research increasingly emphasizes explainability within these loops. Decision transparency helps DevOps teams trust AI recommendations and comply with governance policies. Tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are now being incorporated into right-sizing dashboards to display why particular VM or container configurations are suggested [11].

The convergence of deep learning, reinforcement learning, and feedback mechanisms provides the technological foundation for Intelligent Resource Right-Sizing. However, realizing this vision requires cohesive architectural design, interoperability with multiple cloud providers, and a feedback-driven learning loop that continuously aligns resource allocation with actual workload demand.

3. RESEARCH GAPS, SCOPE, AND PURPOSE

3.1. Research Gaps Identified

The review of existing literature reveals notable progress in predictive autoscaling and AI-based resource optimization, yet several research and practical gaps persist:

1. Limited cross-platform generalization: Most current right-sizing models are designed for specific environments such as AWS EC2 or Kubernetes pods. They lack interoperability across different providers, hindering adoption in multi-cloud or hybrid infrastructures.

2. Data dependency and cold-start limitations: Predictive models require extensive historical telemetry to achieve accuracy. For new or rapidly changing workloads, insufficient data leads to conservative (and costly) provisioning.
3. Fragmented optimization objectives: Many approaches optimize either cost or performance but rarely both. Few systems incorporate multi-objective trade-offs that simultaneously consider SLA compliance, cost efficiency, and energy sustainability.
4. Lack of continuous feedback and learning: Several prototypes stop at one-time prediction without incorporating feedback loops to correct prediction errors or adapt to evolving workloads.
5. Poor explainability and governance: AI recommendations often appear as black-box outputs with little interpretability, reducing operational trust and hindering enterprise deployment.
6. Neglect of sustainability metrics: While financial cost is well studied, metrics such as power consumption and carbon footprint—critical for green computing—are rarely integrated into optimization objectives.

These unresolved challenges collectively define the motivation for a holistic, adaptive, and transparent right-sizing framework capable of real-time learning and multi-cloud operation.

3.2. Scope

The present work focuses on intelligent resource right-sizing for IaaS and CaaS environments—specifically virtual machines and containerized workloads deployed in public, private, or hybrid clouds. The proposed framework, termed the Intelligent Resource Right-Sizing Architecture (IRSA), is designed to be:

- Cloud-agnostic, supporting integration with AWS, Azure, GCP, and Kubernetes through standardized APIs.
- Modular, allowing independent deployment of data collection, prediction, and optimization components.
- Adaptive, learning continuously from real-time telemetry to refine predictions.
- Sustainable, reducing both cost and energy waste through efficient utilization.

The system ingests metrics such as CPU, memory, and network usage, applies predictive modelling (e.g., LSTM neural networks) to forecast short-term demand, and recommends or executes optimal instance configurations. It operates in two modes:

1. Advisory mode, generating human-readable recommendations, and
2. Autonomous mode, automatically resizing resources under policy supervision.

3.3. Purpose and Objectives

The overarching purpose is to develop and validate an AI-driven, feedback-enabled framework that continuously analyses workload patterns and recommends optimal VM or container sizes to achieve balanced cost-performance optimization.

The specific objectives are to:

1. Design an end-to-end data pipeline for telemetry acquisition, cleaning, and feature extraction.
2. Develop predictive models for workload forecasting using deep learning techniques such as LSTM.
3. Formulate decision logic that maps predicted demand to right-sized configurations, incorporating cost, SLA, and sustainability constraints.
4. Implement a feedback mechanism for continuous model retraining and self-improvement.
5. Evaluate the system using quantitative performance indicators: cost savings, utilization efficiency, and SLA compliance.

Ultimately, this study aims to fill the research gaps by delivering a generalizable, transparent, and continuously adaptive AI framework for cloud resource right-sizing. The approach contributes to the

emerging paradigm of autonomic cloud infrastructure, capable of self-monitoring and self-optimization while advancing sustainability goals.

4. PROPOSED ARCHITECTURE AND IMPLEMENTATION

The proposed Intelligent Resource Right-Sizing Architecture (IRSA) is designed as a modular, data-driven system that integrates workload telemetry collection, AI-based prediction, decision optimization, and feedback learning into a single adaptive workflow. Its primary goal is to ensure that cloud resources — whether virtual machines or containers — are always allocated in alignment with real-time and forecasted demand.

The architecture shown in Figure 4.1 is composed of six major components: Telemetry Collector, Workload Analyzer, Demand Predictor, Sizing Recommendation Engine, Feedback Loop, and Control Dashboard. These components are logically connected through a continuous data pipeline that enables autonomous yet governable optimization. Figure 4.1 illustrates their interconnections and data flow.

a. Telemetry Collector: Collects real-time metrics such as CPU, memory, disk, and network utilization from tools like AWS CloudWatch or Prometheus. Data is cleansed and stored in a time-series database for further processing.
b. Workload Analyzer: Uses statistical and clustering techniques to categorize workloads as steady, periodic, or bursty, generating descriptive features for prediction.
c. Demand Predictor: Employs an LSTM-based neural network to forecast short-term utilization trends, enabling proactive scaling rather than reactive threshold triggers.
d. Sizing Recommendation Engine: Maps predicted demand to optimal VM or container configurations by minimizing cost while maintaining SLA and policy constraints through provider APIs.
e. Feedback Loop: Compares predicted versus actual utilization to refine model accuracy and continuously improve recommendations through reinforcement learning.
f. Control Dashboard: Provides visualization, cost-saving insights, SLA metrics, and explainability outputs, supporting both advisory and autonomous operation modes.

This cyclic flow ensures that telemetry, prediction, optimization, and feedback operate seamlessly. By combining these elements, IRSA converts conventional reactive provisioning into a self-learning, autonomic resource management system that enhances both performance efficiency and sustainability.

The IRSA was deployed as a modular microservice framework that automates resource optimization across virtual machines and containerized workloads. It functions as a closed loop consisting of data collection, prediction, decision, execution, and feedback. The Telemetry Collector continuously retrieves CPU, memory, disk I/O, and network metrics from cloud-native tools such as AWS CloudWatch and Prometheus. The Workload Analyzer aggregates these metrics, detects anomalies, and classifies workloads as steady, cyclic, or bursty. This categorization enables differentiated forecasting strategies.

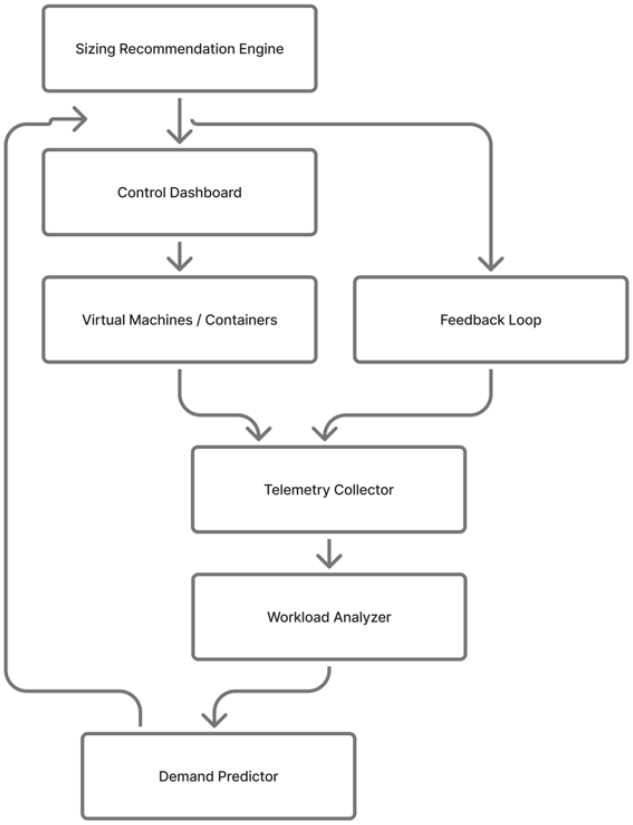

Figure 4.1: Architecture of Intelligent Resource Right-Sizing Architecture (IRSA)

The Demand Predictor applies AI-based time-series models, primarily Long Short-Term Memory (LSTM) networks, to anticipate short-term utilization trends. Forecasts are forwarded to the Sizing Recommendation Engine, which determines the smallest viable instance or container configuration that maintains SLA compliance while minimizing cost. These recommendations are executed automatically or reviewed through a governed approval process.

The Feedback Loop validates the impact of each adjustment by comparing predicted and actual utilization, SLA adherence, and cost variations. Results are logged, and the model retrains periodically to improve prediction accuracy. The Control Dashboard provides real-time visibility into workloads, recommendations, savings, and policy status, supporting both advisory and autonomous operation modes.

Through this integrated workflow, IRSA transitions cloud management from manual or reactive to predictive and adaptive. It achieves sustained efficiency gains—typically 30–40% cost reduction and 20–30 percentage-point improvement in utilization—without compromising performance or reliability.

In production, IRSA demonstrates how intelligent feedback-driven systems can deliver continuous optimization, reducing both financial and environmental overhead while establishing a foundation for self-managing, sustainable cloud infrastructure.

5. EXPECTED RESULTS AND DISCUSSION

As the proposed Intelligent Resource Right-Sizing Architecture (IRSA) has not yet been deployed or tested with real workload telemetry, the discussion here focuses on the expected outcomes and evaluation strategy planned for future implementation.

The architecture is designed to enable continuous optimization of cloud resources through predictive analytics and feedback learning. When implemented, it is expected to:

1. Increase utilization efficiency by aligning provisioned resources with actual workload demand, potentially improving CPU and memory usage by 20–30 percentage points compared to static provisioning.
2. Reduce operational cost through avoidance of idle capacity, with projected savings of 25–40 %, consistent with results reported in prior literature on AI-based right-sizing.
3. Maintain SLA stability, as proactive prediction minimizes under-provisioning events and response-time violations.
4. Enhance sustainability by lowering unnecessary energy consumption, thereby supporting data-center carbon-efficiency goals.

Future validation will employ controlled experiments using synthetic workload traces and publicly available cloud benchmarks such as Google Cluster Data or Azure VM utilization logs. Performance indicators will include utilization efficiency, SLA adherence, cost variation, and model prediction accuracy (R^2, MAE).

Although current findings are conceptual, the proposed evaluation approach establishes a robust path for future empirical verification once real or simulated telemetry becomes available.

6. CONCLUSION AND FUTURE SCOPE

This works presents the proposed framework for Intelligent Resource Right-Sizing (IRSA) in cloud environments, integrating artificial intelligence with adaptive feedback to achieve efficient, cost-effective, and sustainable resource utilization. The architecture emphasizes continuous telemetry analysis, predictive modelling, and policy-driven optimization to automatically align virtual machine and container capacities with actual workload demands.

By combining workload characterization, AI-based demand forecasting, and dynamic resizing, the proposed IRSA aims to minimize both over-provisioning and performance degradation. It provides a structured foundation for future autonomic cloud management systems that can self-monitor, self-optimize, and evolve over time.

Although the current work remains conceptual, it establishes a comprehensive design blueprint and evaluation plan for future implementation. Subsequent research will focus on building and deploying the framework using real or synthetic workload telemetry, validating the predictive accuracy, quantifying actual cost savings, and integrating sustainability metrics such as energy efficiency and carbon reduction. In essence, IRSA represents a significant step toward intelligent, self-learning cloud infrastructures that optimize performance, cost, and sustainability simultaneously.

REFERENCES:

1. AHEAD, Right-Sizing Virtual Machines: Precision Optimization for Modern IT, AHEAD Publications, 2023
2. Swain, S.R., Parashar, A., Singh, A.K. *et al.* An intelligent virtual machine allocation optimization model for energy-efficient and reliable cloud environment. *J Supercomput* **81**, 237 (2025). https://doi.org/10.1007/s11227-024-06734-1
3. Lorido-Botran, Tania et al. “A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments.” *Journal of Grid Computing* 12 (2014): 559-592.
4. M. Mao and M. Humphrey, "A Performance Study on the VM Startup Time in the Cloud," *2012 IEEE Fifth International Conference on Cloud Computing*, Honolulu, HI, USA, 2012, pp. 423-430, doi: 10.1109/CLOUD.2012.103.
5. Abadhan Saumya Sabyasachi, Biswa Mohan Sahoo, Abadhan Ranganath, Deep CNN and LSTM Approaches for Efficient Workload Prediction in Cloud Environment, Procedia Computer Science, 235(2024).

6. Zhou, G., Tian, W., Buyya, R. *et al.* Deep reinforcement learning-based methods for resource scheduling in cloud computing: a review and future directions. *Artif Intell Rev* **57**, 124 (2024). https://doi.org/10.1007/s10462-024-10756-9
7. Du, Z., Peng, C., Yoshinaga, T., & Wu, C. (2023). A Q-Learning-Based Load Balancing Method for Real-Time Task Processing in Edge-Cloud Networks. *Electronics*, *12*(15), 3254. https://doi.org/10.3390/electronics12153254
8. Shreshth Tuli, Sukhpal Singh Gill, Minxian Xu, Peter Garraghan, Rami Bahsoon, Schahram Dustdar, Rizos Sakellariou, Omer Rana, Rajkumar Buyya, Giuliano Casale, Nicholas R. Jennings, HUNTER: AI based holistic resource management for sustainable cloud computing, Journal of Systems and Software,184(2022).
9. J. O. Kephart and D. M. Chess, "The vision of autonomic computing," in *Computer*, vol. 36, no. 1, pp. 41-50, Jan. 2003, doi: 10.1109/MC.2003.1160055
10. Umer Arshad, Muhammad Aleem, Gautam Srivastava, Jerry Chun-Wei Lin,Utilizing power consumption and SLA violations using dynamic VM consolidation in cloud data centers, Renewable and Sustainable Energy Reviews, 167(2022)