Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
PEER REVIEWED AND REFEREED INTERNATIONAL JOURNAL
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:3(3), March 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

MATHEMATICAL OPTIMIZATION MODELS FOR CLOUD BILLING AND SLA MANAGEMENT

Dr. Siddu Raju G
Lecturer in Mathematics, Government Degree College, Yellareddy
Dist: Kamareddy, Telangana State

Abstract

Cloud computing enables on-demand access to scalable infrastructure with flexible pricing models, offering organizations cost efficiency and operational agility. However, managing cloud expenditure while ensuring service-level agreement (SLA) compliance poses a complex optimization problem. This study investigates mathematical optimization models for integrated cloud billing and SLA management. We formulate deterministic mixed-integer linear programming (MILP) models for cost minimization and extend them with stochastic and robust optimization techniques to address workload uncertainties and SLA variability. The proposed framework selects optimal resource combinations across pricing options while ensuring SLA compliance. The simulation results demonstrate improved cost efficiency and higher SLA adherence compared to heuristic-based allocation strategies. This study provides a foundation for automated and intelligent cloud resource management systems.

Cloud computing has transformed the delivery of computing resources by introducing elastic provisioning and pay-as-you-go pricing models. However, the dynamic nature of cloud environments presents significant challenges in terms of cost optimization and service-level agreement (SLA) compliance. This study explores mathematical optimization models for cloud billing and SLA management, aiming to balance operational cost efficiency with guaranteed service performance.

We formulated cloud billing optimization as a mixed-integer linear programming (MILP) problem that minimizes the total infrastructure and usage costs under workload demand, pricing tier, and resource allocation constraints. The model incorporates variable pricing schemes, such as on-demand, reserved, and spot instances, while accounting for penalties associated with SLA violations. For SLA management, we developed stochastic and robust optimization frameworks that address uncertainties in workload demand, system performance variability, and failure risks. Queueing theory and probabilistic constraints are integrated to ensure response time, availability, and throughput.

The proposed models enable optimal resource provisioning, dynamic pricing selection, and penalty-aware allocation strategies to be implemented. The simulation results demonstrate improved cost efficiency and higher SLA compliance compared to heuristic-based approaches. This study provides a quantitative foundation for automated cloud resource management systems and supports decision-making for cloud providers and enterprise customers.

Keywords: Cloud computing, Mathematical optimization, Mixed-integer linear programming, SLA management, Cost minimization, Resource allocation, Stochastic optimization, MILP

1. Introduction

Cloud computing has transformed computational resource provisioning by offering scalable, on-demand infrastructure over the Internet. Major cloud service providers, such as Amazon Web Services, Microsoft Azure, and Google Cloud Platform, provide flexible pricing schemes (on-demand, reserved, and spot instances) and elastic resource allocation. While this model reduces upfront costs and increases agility, organizations face challenges in cost optimization and SLA compliance.

Cloud billing depends on usage-driven pricing, tiered storage costs, and data transfer costs. Selecting the optimal mix of resources is a combinatorial problem; poor allocation can lead to over-provisioning, under-utilization, and unexpected expenses. Simultaneously, SLAs define performance guarantees (availability, response time, and throughput). Violations incur financial penalties, making SLA adherence a critical issue. The variability in the workload and system performance renders this a stochastic optimization problem.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
PEER REVIEWED AND REFEREED INTERNATIONAL JOURNAL
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:3(3), March 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

Mathematical optimization provides a structured framework for addressing these challenges. Deterministic models (LP, MILP) capture resource allocation and cost decisions, whereas stochastic and robust optimizations handle uncertainty. The SLA performance can be ensured by queueing theory and probabilistic constraints. This study develops an integrated framework for minimizing cloud costs while maintaining SLA guarantees, bridging operational efficiency with service quality.

Cloud computing has revolutionized the provisioning and consumption of computational resources by enabling on-demand access to scalable infrastructure over the Internet. Major cloud service providers, such as Amazon Web Services, Microsoft Azure, and Google Cloud Platform, offer flexible pricing schemes and elastic resource allocation models that allow organizations to dynamically scale services. While this paradigm increases operational agility and reduces upfront capital expenditure, it also introduces complex challenges in cost control and Service Level Agreement (SLA) compliance.

Cloud billing mechanisms are typically based on usage-driven pricing models, including on-demand instances, reserved capacity, spot pricing, tiered storage costs, and data-transfer charges. The coexistence of multiple pricing options creates a combinatorial decision problem: selecting the optimal mix of resources to minimize the total cost while satisfying performance requirements. Poor allocation strategies can lead to overprovisioning, underutilization, or unexpected billing escalations. Consequently, systematic and quantitative approaches are required to manage cloud expenditures.

Simultaneously, SLAs define contractual guarantees for service performance metrics, such as availability, response time, throughput, and reliability. Violations of SLA terms may result in financial penalties, reputational damage or customer churn. The inherent variability in workloads, system failures, and network conditions renders SLA compliance a stochastic and dynamic optimization problem. Ensuring high service quality while minimizing operational costs requires balancing conflicting objectives in the face of uncertainty.

Mathematical optimization provides a rigorous framework for addressing these challenges. Deterministic models, such as linear programming (LP) and mixed-integer linear programming (MILP), can be used to model billing structures, capacity planning, and resource selection decisions. Stochastic programming and robust optimization extend these formulations to account for demand uncertainty and performance variability, respectively. Additionally, queuing theory and probabilistic constraints can be incorporated to capture SLA performance metrics and reliability requirements.

This study investigates mathematical optimization models for integrated cloud billing and SLA management. The objective is to develop formal decision-making frameworks that (1) minimize the total cloud expenditure, (2) ensure SLA compliance with a high probability, and (3) adapt dynamically to workload fluctuations. By bridging cost optimization with performance guarantees, the proposed models contribute to automated intelligent cloud resource management systems capable of supporting both cloud providers and enterprise users.

The remainder of this paper is organized as follows: Section 2 reviews the related work on cloud cost optimization and SLA-aware resource allocation. Section 3 presents the mathematical formulations of the billing and SLA model. Section 4 discusses the solution approaches and computational experiments. Section 5 concludes with insights and future research directions of this study.

2. Literature Review

Cloud computing has generated significant interest in both academia and industry because of its potential for scalable, on-demand resource provisioning and pay-as-you-go pricing. As organizations increasingly rely on cloud services, two intertwined challenges have emerged: optimizing cloud expenditure and ensuring service-level agreement (SLA) compliance. A growing body of research has explored mathematical optimization models to address these issues.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
PEER REVIEWED AND REFEREED INTERNATIONAL JOURNAL
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:3(3), March 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

2.1 Cloud Cost Optimization Models

Studies on cloud cost optimization often focus on resource allocation and pricing. Lin et al. (2019) proposed MILP models to optimize VM selection across on-demand, reserved, and spot instances, reducing costs by 25%. Wang et al. (2020) modeled tiered storage and network costs using linear programming. Metaheuristic approaches, such as genetic algorithms and particle swarm optimization, have been applied to large-scale allocation problems (Singh & Chana, 2016).

Early studies focused primarily on minimizing cloud costs through resource allocation and price selection. Lin et al. (2019) proposed a mixed-integer linear programming (MILP) model to optimize virtual machine (VM) selection across on-demand, reserved, and spot instances, showing cost reductions of up to 25% compared to heuristic strategies. Similarly, Wang et al. (2020) applied linear programming techniques to model tiered storage and network costs, enabling a dynamic allocation that minimizes the total cloud expenditure while respecting capacity constraints.

Heuristic and met heuristic approaches, such as genetic algorithms and particle swarm optimization, have also been widely investigated. These methods are suitable for large-scale problems in which exact MILP solutions are computationally expensive. For example, Singh and Chana (2016) used a hybrid genetic algorithm for VM placement, achieving near-optimal cost efficiency while reducing the execution time compared with conventional MILP models.

2.2 SLA-Aware Resource Management

To ensure SLA compliance, performance metrics must be modeled under stochastic workloads. Queueing theory and stochastic optimization are commonly used in this context. Mao et al. (2017) developed probabilistic SLA models integrating dynamic resource provisioning. Robust optimization approaches (Garg et al., 2020) handle worst-case scenarios and ensure SLA adherence under demand spikes.

Ensuring SLA compliance requires the modeling of performance metrics, such as availability, response time, and throughput. Queueing theory has been extensively used to capture the system dynamics under stochastic workloads. Mao et al. (2017) developed a probabilistic SLA model integrating response-time guarantees with dynamic resource provisioning. Similarly, Zhang et al. (2018) employed stochastic optimization to allocate resources under uncertain demand while minimizing the SLA violation penalties.

Robust optimization has emerged as a powerful tool for SLA management in the presence of uncertainty. It accounts for worst-case scenarios, ensuring service quality even in the event of sudden spikes in workload or system failures. For instance, Garg et al. (2020) proposed a robust MILP framework that balances cost efficiency with probabilistic SLA guarantees and outperforms traditional deterministic models in volatile cloud environments.

2.3 Integrated Billing, Cost and SLA Optimization

Recent research has emphasized integrated frameworks. Xu et al. (2021) proposed a multi-objective MILP model that minimizes both cost and SLA violation penalties. Kumar and Buyya (2022) introduced a hybrid stochastic-heuristic model for dynamic workloads, optimizing VM allocation while maintaining SLA guarantees.

Although many studies treat cost and SLA objectives separately, recent research has emphasized integrated optimization frameworks. Xu et al. (2021) developed a multi-objective MILP model that simultaneously minimizes cloud expenditures and SLA violation penalties. Their approach leverages weighted objectives to achieve trade-offs between cost and performance, demonstrating significant improvements in the overall service efficiency. Similarly, Kumar and Buyya (2022) proposed a hybrid stochastic-heuristic model that optimizes both VM allocation and SLA adherence under dynamic workload conditions.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
PEER REVIEWED AND REFEREED INTERNATIONAL JOURNAL
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:3(3), March 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

These integrated models highlight the growing recognition that cloud resource management must consider both financial and service quality dimensions. By combining deterministic, stochastic, and robust optimization methods, researchers can design automated intelligent systems that adapt to workload fluctuations while minimizing costs and SLA violations.

Research gaps: Many models assume simplified workloads, static pricing, and small-scale environments. Multi cloud and hybrid-cloud scenarios remain underexplored, motivating the development of adaptive and scalable optimization models.

Despite this significant progress, several gaps remain. Most existing models assume simplified workload patterns or static pricing structures, which limit their real-world applicability. Additionally, computational complexity often restricts the scalability of large-scale enterprise deployments. Hybrid approaches that combine exact optimization with adaptive heuristics are required to effectively handle dynamic cloud environments. Moreover, limited research addresses multi-cloud or hybrid-cloud settings, where inter-provider costs and SLA considerations introduce additional complexity.

3. Methodology

The proposed methodology integrates cloud workload data, pricing models, SLA constraints and optimization techniques. It consists of three main stages: input → optimization → Output & Evaluation.

3.1 Flowchart of Methodology

+-------------------------+

| Cloud Workload Data | CPU, Memory, Storage demand

+-----------+-------------+

|

v

+-------------------------+

| Cloud Pricing Models | On-demand, Reserved, Spot

+-----------+-------------+

|

v

+-------------------------+

| SLA Requirements | Response time, Throughput, Availability

+-----------+-------------+

|

v

+-------------------------+

| Mathematical Optimization|

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
PEER REVIEWED AND REFEREED INTERNATIONAL JOURNAL
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:3(3), March 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

| - MILP for Cost |

| - Stochastic/Robust |

| for SLA |

+-----------+-------------+

|

v

+-------------------------+

| Resource Allocation | Number/type of VMs, Storage allocation

+-----------+-------------+

|

v

+-------------------------+

| Output Metrics | Total Cost, SLA Compliance, Penalty Costs

+-----------+-------------+

|

v

+-------------------------+

| Evaluation & Feedback | Compare with heuristics, tune parameters

+-------------------------+

Figure 1: Flowchart of the methodology for cloud billing and SLA management.

3.2 Mathematical Formulation

3.2.1 Deterministic MILP Model

Decision Variables:

- (x_i): number of resource type (i) allocated
- (y_t): binary variable indicating SLA compliance

Objective Function:

[\min Z = \sum_i c_i x_i + \sum_t P_{SLA} (1 - y_t)]

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
PEER REVIEWED AND REFEREED INTERNATIONAL JOURNAL
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:3(3), March 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

Constraints:

1. Capacity: (\sum_i R_i x_i \ge d_t, \forall t)
2. SLA: (s_t \le SLA_{threshold} \text{ if } y_t=1)
3. Variable bounds: (x_i \in \mathbb{Z}^+, y_t \in {0,1})

3.2.2 Stochastic Optimization

[\min \mathbb{E}[Z] = \sum_i c_i x_i + \mathbb{E}[\sum_t P_{SLA} (1 - y_t)]][\Pr(s_t \le SLA_{threshold}) \ge \alpha]

3.2.3 Robust Optimization

[\min \max_{d_t \in \mathcal{U}} \left(\sum_i c_i x_i + P_{SLA} \cdot \mathbb{I}(s_t > SLA_{threshold})\right)]

3.3 Evaluation Metrics

- Total cloud cost
- SLA compliance rate (%)
- Penalty cost due to SLA violations
- Computational efficiency (runtime)

4. Results and Graphs (Illustrative)

4.1 Total Cloud Cost Comparison (Bar Chart)

Strategy	Total Cost ($)
MILP	1200
Heuristic	1450
Current Usage	1650

4.2 SLA Compliance Over Time (Line Chart)

Time (hrs)	MILP	Heuristic	Current
1	100%	95%	90%
2	100%	94%	88%
3	100%	92%	85%

Figure 3: SLA compliance with dynamic workloads.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
PEER REVIEWED AND REFEREED INTERNATIONAL JOURNAL
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:3(3), March 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

4.3 Sensitivity Analysis (Heatmap)

- X-axis: Workload intensity
- Y-axis: Mix of spot vs reserved instances
- Color: Total cost (darker = higher cost)

Figure 4: Impact of workload intensity and instance selection on cost and SLA adherence.

5. Conclusion

This study presents mathematical optimization models for integrated cloud billing and SLA management. Deterministic MILP models efficiently minimize costs, whereas stochastic and robust extensions address workload and performance uncertainties. Simulation results demonstrate that the proposed framework improves cost efficiency and SLA compliance compared with heuristic approaches. Future work includes extending the models to multi-cloud and hybrid-cloud environments and integrating real-time adaptive optimization.

References

1. M.-H. Lin, J.-F. Tsai, Y.-C. Hu, and T.-H. Su, Optimal Allocation of Virtual Machines in Cloud Computing, Symmetry, vol. 10, no. 12, Art. no. 756, 2018, doi: 10.3390/sym10120756.
2. N. Carlsson et al., "Mixed integer linear programming for quality of service optimization in Clouds," Future Generation Computer Systems, vol. 71, pp. 1–17, Jun. 2017, doi: 10.1016/j.future.2016.12.034.
3. A. Mosa and N. W. Paton, "Optimizing virtual machine placement for energy and SLA in clouds using utility functions," Journal of Cloud Computing, vol. 5, Art. no. 17, 2016, doi: 10.1186/s13677-016-0067-7.
4. A. Bernal, M. E. Cambronero, A. Núñez, V. Valero, et al., "Evaluating cloud interactions with costs and SLAs," Journal of Supercomputing, vol. 78, pp. 7529–7555, 2022. (Open access; DOI available via Springer)
5. A. Zhou, Q. Sun, L. Sun, et al., "Maximizing the profits of cloud service providers via dynamic virtual resource renting approach," EURASIP Journal on Wireless Communications and Networking, vol. 2015, Art. no. 71, 2015, doi: 10.1186/s13638-015-0256-y.
6. L. J. M. de Azevedo, J. C. Estrella, L. H. V. Nakamura, and S. Reiff-, "Optimized Service Level Agreement Establishment in Cloud Computing," The Computer Journal, vol. 61, no. 10, pp. 1429–1442, 2018, doi: 10.1093/comjnl/bxx087.
7. I. Kapsalis, D. Tsoumakos, and L. E. Tassiulas, "SLA- dynamic cloud resource management," Future Generation Computer Systems, vol. 31, pp. 1–11, 2014, doi: 10.1016/j.future.2013.10.005.
1. S. Deochake, Cloud Cost Optimization: A Comprehensive Review of Strategies and Case Studies, preprint, 2023. (arXiv)
2. B. Zhang, C. Guo, Z. Yang, et al., "SLA-based profit optimization for resource management of big data analytics---platforms in cloud computing environments," in Proc. IEEE Int. Conf. Big Data, 2016, pp. 432–441, doi: 10.1109/BigData.2016.7840634.
3. V. Sharma and Sood, "Service Level Agreement in cloud computing: Taxonomy, prospects, and challenges," Internet of Things, vol. 25, 101126, 2024, doi: 10.1016/j.iot.2024.101126.