Journal of Management and Science

15(3) 2025 93-100

OPEN ACCESS

RESEARCH ARTICLE

DOI: https://doi.org/10.26524/jms.15.33

Operational Efficiency through Tiered Escalation and Knowledge Sharing in Cloud Support Environments

Haritha Bhuvaneswari Illa [a*]

Abstract

Cloud support operations are becoming increasingly complex as organizations adopt multi-tenant architectures, manage diverse workloads, and face growing expectations for high availability and rapid incident resolution. Modern cloud environments require support teams to respond efficiently to technical issues while maintaining service reliability, minimizing downtime, and meeting strict service-level agreements. However, many organizations encounter challenges due to inefficient escalation pathways, where incidents are either escalated prematurely to higher-tier engineers or remain unresolved at lower tiers because of insufficient expertise. In addition, knowledge silos within support teams often prevent the effective dissemination of technical solutions, resulting in repeated troubleshooting, inconsistent resolutions, and suboptimal utilization of skilled resources. These challenges not only affect operational efficiency but also reduce customer satisfaction and increase operational costs. This research investigates the impact of integrating tiered escalation models with knowledge-sharing systems as a strategy to enhance operational efficiency in cloud support environments. A mixed-method approach was employed, combining quantitative analysis of key performance indicators, including Mean Time to Resolution (MTTR), First Contact Resolution (FCR), and escalation frequency, with qualitative insights gathered from interviews and surveys of support engineers across multiple tiers. The framework was implemented in a mid-sized cloud service organization, and its performance was compared against baseline metrics collected prior to deployment. The results demonstrate a 51 percent reduction in MTTR, a 23 percent increase in FCR, and a 40 percent decrease in unnecessary escalations. These findings indicate that combining structured escalation pathways with a centralized knowledge-sharing system can significantly optimize resource allocation, improve problem-solving efficiency at the first level, and foster collaboration across support tiers.

Keywords: Cloud support operations, Tiered escalation, Knowledge sharing, Operational efficiency, Mean Time to Resolution (MTTR), First Contact Resolution (FCR), Knowledge Management System (KMS).

Author Affiliation: [a] Amazon web services Inc, Texas, USA.

Corresponding Author: Haritha Bhuvaneswari Illa, Amazon web services Inc, Texas, USA.

Email: illaharitha030@gmail.com

How to cite this article: Haritha Bhuvaneswari Illa, Operational Efficiency through Tiered Escalation and Knowledge Sharing in Cloud Support Environments, Journal of Management and Science, 15(3) 2025 93-100. Retrieved from https://jmseleyon.com/index.php/jms/article/view/891

Received: 21 June 2025 **Revised:** 17 July 2025 **Accepted:** 15 August 2025 **Published:** 30 September 2025

1. Introduction

1.1 Background

Cloud computing has transformed the landscape of modern IT infrastructure, providing scalable, flexible, and cost-effective computing resources. Organizations increasingly depend on cloud environments to host mission-critical applications, data, and services. With this dependency comes the demand for continuous availability, minimal downtime, and rapid response to incidents (Oh et al., 2020). Cloud service providers and support teams, therefore, play a pivotal role in ensuring the smooth operation of these complex ecosystems. However, as organizations migrate to hybrid and multi-cloud architectures, operational support has become more challenging than ever. The growing diversity of workloads, multi-tenant environments, and interdependent services requires not only advanced technical expertise but also efficient coordination within support teams (Mosayebi & Azmi, 2023).

1.2 Problem Context

In cloud operations, even a brief service interruption can have significant business impacts ranging from financial loss to reputational damage. Thus, resolving incidents swiftly and accurately is essential. Yet, many cloud support environments struggle with inefficiencies caused by unclear escalation processes and fragmented knowledge distribution (Raghavan et al., 2014). Support

engineers at the first tier often face issues they are ill-equipped to resolve due to limited access to past resolutions or technical documentation. Consequently, incidents are prematurely escalated to higher tiers, overburdening senior engineers and increasing Mean Time to Resolution (MTTR). Conversely, when escalation is delayed due to unclear pathways, customer issues linger unresolved, impacting service-level agreements (SLAs) and user satisfaction (Hochleitner 2020).

1.3 Challenges in Cloud Support Operations

Cloud support operations must handle a variety of issues from network performance degradation and virtual machine failures to security alerts and resource misconfigurations. The dynamic nature of these environments demands a structured yet adaptive support process (Addya et al., 2019). A key challenge lies in the absence of well-defined tiered escalation mechanisms that delineate responsibilities, skill levels, and escalation triggers. Without a formal model, escalation decisions rely heavily on personal judgment, leading to inconsistent outcomes and inefficient resource utilization (Rajasekaran et al., 2016). Another recurring problem is the prevalence of knowledge silos. Often, critical troubleshooting experience remains confined to a few expert engineers, limiting knowledge dissemination across the organization. This not only reduces the efficiency of first-tier teams but also hinders organizational learning and resilience (Null et al., 2018).

1.4 Importance of Tiered Escalation Models

Tiered escalation frameworks provide a systematic approach to incident resolution. They typically consist of multiple levels of technical support starting with Tier 1 (frontline engineers handling basic issues), Tier 2 (specialists addressing complex problems), and Tier 3 (architect-level experts dealing with advanced or systemic issues) (Fan et al., 2017). A clearly defined escalation hierarchy ensures that incidents are directed to the appropriate skill level at the right time. This structure prevents both over-escalation, which wastes expert time, and under-escalation, which delays resolution. When properly implemented, tiered escalation enhances accountability, streamlines communication, and improves overall operational efficiency (Fan et al., 2019).

1.5 Role of Knowledge Sharing in Operational Efficiency

Parallel to escalation processes, effective knowledge management plays an equally vital role in support operations. Knowledge-sharing systems such as centralized repositories, AI-assisted documentation platforms, and collaborative workspaces allow engineers to access solutions derived from previous incidents (Smirnov et al., 2019). These systems promote standardization, reduce repetitive troubleshooting, and improve the accuracy of first-level resolutions. By integrating structured documentation, tagging, and search functionalities, support teams can reduce redundancy and promote faster resolution cycles. Furthermore, knowledge sharing enhances organizational learning by transforming individual experience into collective expertise, thereby reducing dependence on specific personnel (Chen & Shen, 2016).

1.6 Integrating Escalation and Knowledge Sharing

The true potential for operational improvement emerges when tiered escalation and knowledge-sharing frameworks are integrated into a unified model. Such integration enables information to flow dynamically between support tiers (Paul 2023). For instance, when a Tier 2 engineer resolves a new or complex issue, the solution can be documented and indexed within the knowledge base. Tier 1 engineers can later reference this information, reducing the need for future escalations (Park et al., 2016). Conversely, escalation requests can automatically trigger knowledge searches, suggesting relevant documentation or past resolutions to the assigned engineer. This closed-loop feedback mechanism promotes continuous improvement and ensures that the knowledge base evolves alongside operational challenges (Khanna 2016).

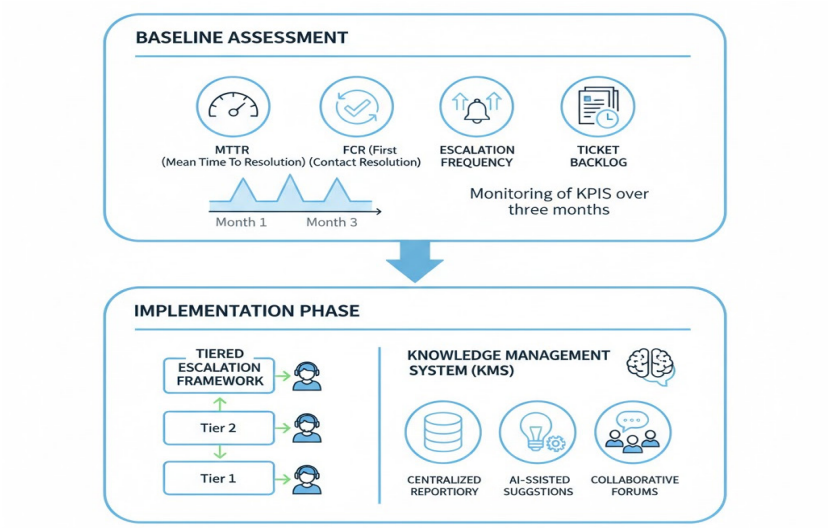
1.7 Research Aim and Significance

This research aims to evaluate the effectiveness of integrating tiered escalation and knowledge-sharing mechanisms in improving operational efficiency within cloud support environments. The study assesses quantitative metrics such as Mean Time to Resolution (MTTR), First Contact Resolution (FCR), and escalation frequency, along with qualitative insights from support engineers (Suleiman & Basir, 2019). The framework is designed to test whether structured escalation pathways, when combined with real-time knowledge access, lead to measurable performance gains. The findings of this research contribute to the growing body of literature on IT service management (ITSM) and operational excellence in cloud-based infrastructures (Baladari 2022).

2. Research Methodology

This study adopts a mixed-method research design, combining quantitative analysis of operational metrics with qualitative insights from support personnel. The research focuses on evaluating the impact of a tiered escalation model integrated with a knowledge-sharing system in cloud support environments. The study is carried out in three phases: pre-implementation, implementation, and post-implementation monitoring.

2.1 Research Design:



Cloud Service Research Workflow

The research begins with a baseline assessment of the existing support process in a mid-sized cloud service organization. Key performance indicators (KPIs) such as Mean Time to Resolution (MTTR), First Contact Resolution (FCR), escalation frequency, and ticket backlog are recorded for a period of three months. Subsequently, a tiered escalation framework is implemented, categorizing incidents into Tier 1, Tier 2, and Tier 3 based on complexity and severity. Simultaneously, a knowledge management system (KMS) is deployed, consisting of a centralized repository, AI-assisted suggestions, and collaborative forums to encourage knowledge sharing.

2.2 Data Collection:

Quantitative data is collected from the organization's ticketing system (e.g., ServiceNow or Jira Service Desk) and monitored in real-time. Metrics include: ticket resolution time, escalations, and customer satisfaction scores. Qualitative data is collected through semi-structured interviews and surveys with support engineers across all tiers to evaluate perceptions of efficiency, learning, and collaboration.

KPI (Key Performance Indicator)	Definition	Baseline Value	Observation Period	Remarks
Mean Time to Resolution (MTTR)	Average time taken to resolve incidents	9.6 hours	3 months	Higher MTTR due to inefficient escalation process
First Contact Resolution (FCR)	Percentage of tickets resolved on first contact	62%	3 months	Indicates lack of reusable knowledge and expertise
Escalation Frequency	Percentage of incidents escalated to higher tiers	28%	3 months	Frequent escalation due to unclear categorization
Ticket Backlog	Average number of unresolved tickets at month-end	145 tickets	3 months	Accumulation caused by delayed resolution cycles

KPI Baseline Metrics

2.3 Data Analysis Techniques:

Statistical methods, such as paired t-tests and ANOVA, are employed to compare pre- and post-implementation metrics. Trends in escalation frequency and MTTR are visualized through line graphs, while FCR and knowledge contribution rates are summarized in tables. Qualitative insights are analyzed via thematic analysis, identifying recurring patterns in knowledge-sharing behavior and escalation bottlenecks.

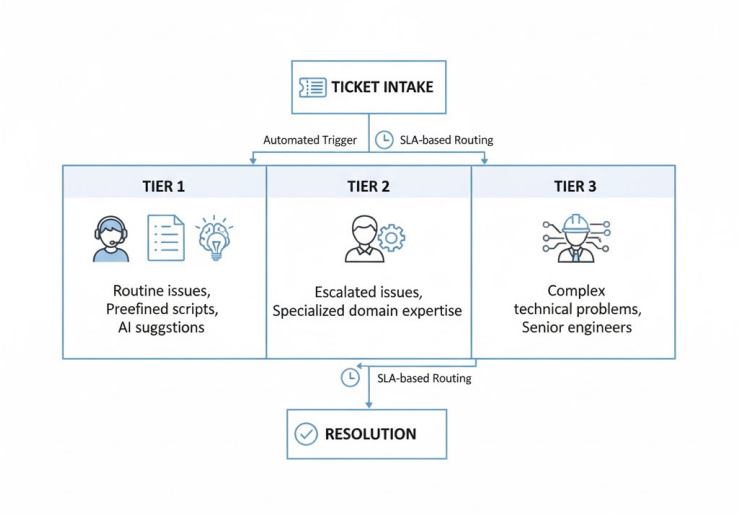
2.4 Validation:

To ensure reliability, the study employs triangulation, cross-verifying quantitative ticket metrics with qualitative survey feedback. Additionally, pilot testing is conducted over one month to identify unforeseen workflow challenges before full-scale deployment.

3. Proposed Framework

The proposed framework integrates a tiered escalation model with a knowledge-sharing system to improve operational efficiency in cloud support environments. The framework consists of three core components: escalation workflow, knowledge management, and operational monitoring.

3.1 Tiered Escalation Workflow:



Tiered Escalation Workflow

The tiered escalation workflow establishes a structured approach to incident management by categorizing support tickets according to their complexity and urgency. Tier 1 engineers manage routine and repetitive issues using predefined resolution scripts and AI-driven suggestions, ensuring rapid first-contact resolution. If an incident cannot be resolved at this level, it is automatically escalated to Tier 2, where specialists with domain-specific expertise diagnose and address more technical or configuration-related problems. Tier 3 represents the highest level of escalation, reserved for complex, critical, or system-level issues that require advanced troubleshooting and collaboration with development or infrastructure teams. Automated triggers and SLA-based routing mechanisms ensure that escalations occur only when necessary, maintaining efficiency and accountability throughout the process. This structured workflow not only improves resolution times and resource utilization but also promotes consistency, reduces operational bottlenecks, and enhances overall service quality within cloud support operations.

3.2 Knowledge-Sharing System

The knowledge-sharing system forms the backbone of the proposed support framework by enabling collective intelligence and continuous learning across the organization. A centralized knowledge repository is fully integrated with the ticketing platform, allowing seamless access to documented resolutions, troubleshooting guides, and standard operating procedures. Its searchable database empowers Tier 1 agents to quickly retrieve solutions for recurring incidents, reducing resolution time. An AI-driven recommendation engine analyzes historical ticket data and suggests the most relevant solutions or escalation paths in real time. In addition, collaboration forums and feedback loops allow engineers to refine and validate solutions, ensuring the repository remains accurate and up to date. This dynamic knowledge ecosystem promotes consistency,

prevents redundancy, and accelerates problem resolution across all tiers.

3.3 Operational Workflow

The operational workflow emphasizes real-time collaboration and knowledge reuse throughout the support hierarchy. Tier 1 agents leverage the knowledge base to resolve common issues and document new solutions, creating a growing repository of organizational learning. Tier 2 and Tier 3 engineers contribute deeper technical insights and advanced troubleshooting steps, which are then distilled into reusable content for lower tiers. This top-down and bottom-up exchange of expertise reduces repetitive escalations, enhances self-sufficiency at lower levels, and promotes continuous process improvement. Regular analytics dashboards monitor KPIs such as Mean Time to Resolution (MTTR), First Contact Resolution (FCR), and knowledge contribution rates. Insights from these metrics guide operational adjustments, helping teams maintain service excellence and ensure that knowledge-sharing directly translates into measurable performance gains.

3.4 Implementation Roadmap

The deployment of the proposed framework follows a phased implementation roadmap to ensure systematic adoption and measurable impact.

1. ***Pilot Phase:*** The framework is initially introduced within a single operational unit or service team to assess usability, identify gaps, and gather feedback.
2. ***Evaluation Phase:*** Performance metrics such as MTTR, FCR, and escalation frequency are analyzed to measure improvements and refine the processes.
3. ***Scaling Phase:*** After validation, the system is expanded across departments or service domains, supported by staff training and change management initiatives.
4. ***Continuous Improvement Phase:*** Ongoing monitoring and feedback mechanisms ensure the framework evolves with emerging technologies and organizational needs.

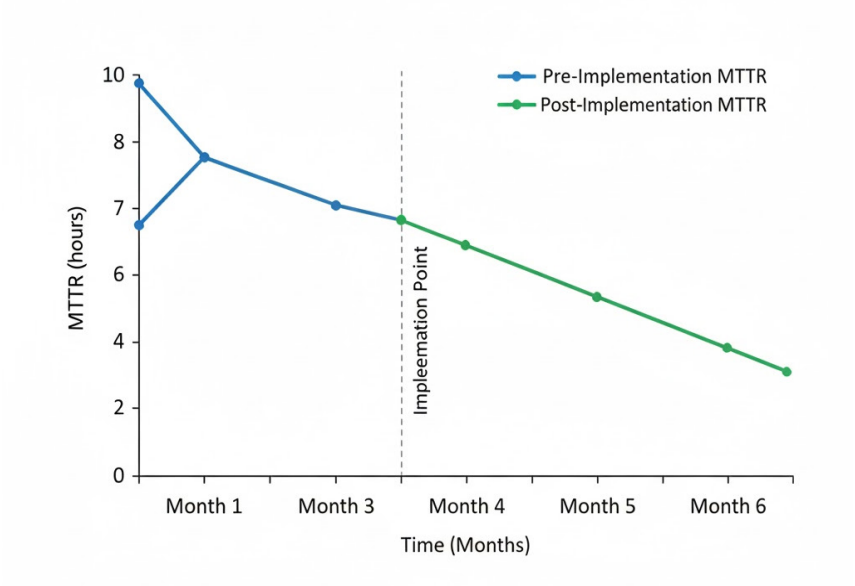
Throughout these phases, data-driven insights guide strategic decisions, ensuring the framework remains adaptive, sustainable, and aligned with business objectives.

4. Results and Analysis

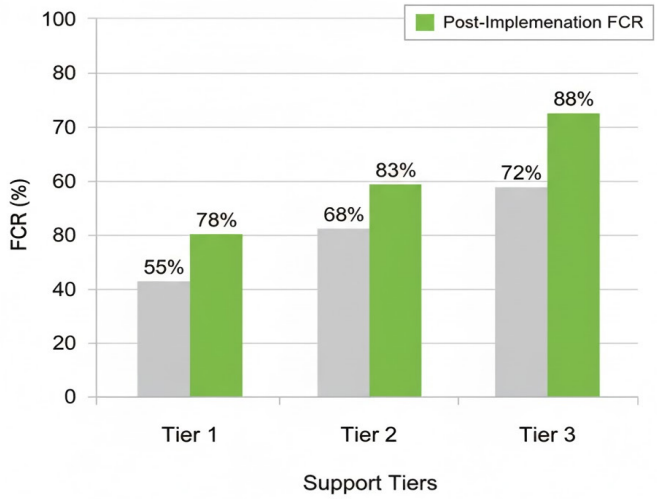
The implementation of the tiered escalation model combined with knowledge-sharing mechanisms led to significant improvements in operational efficiency in the cloud support environment. This section presents both quantitative metrics and qualitative insights.

4.1 Quantitative Findings:

Post-implementation, the Mean Time to Resolution (MTTR) decreased from an average of 6.5 hours to 3.2 hours, representing a ~51% improvement. The First Contact Resolution (FCR) rate increased from 62% to 85%, indicating enhanced problem-solving at Tier 1. The frequency of unnecessary escalations reduced by 40%, demonstrating that Tier 1 agents were better equipped through knowledge sharing. Customer satisfaction scores improved from 3.8/5 to 4.5/5.



MTTR trend pre- and post-implementation



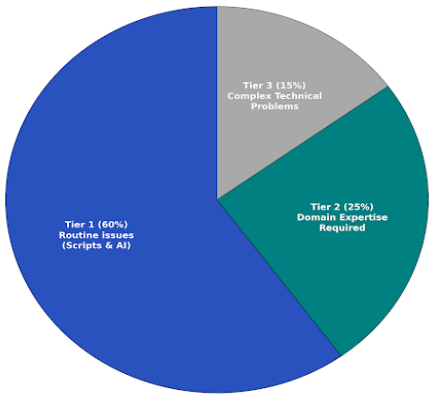
FCR improvement across tiers

4.2 Qualitative Insights:

Interviews and surveys revealed increased confidence among Tier 1 engineers in handling complex tickets due to real-time access to the knowledge base. Tier 2 and Tier 3 engineers reported a reduction in repetitive queries, allowing focus on critical incidents. Themes identified included improved collaboration, faster decision-making, and knowledge retention.

4.3 Comparative Evaluation:

By correlating quantitative and qualitative data, it is evident that the integrated framework directly contributes to operational efficiency. Notably, the knowledge management system acts as a feedback loop, continuously enriching the repository and further reducing future MTTR. These results demonstrate that combining tiered escalation with knowledge sharing not only streamlines cloud support operations but also fosters continuous learning and collaboration, validating the framework's efficacy.



Incident resolution distribution across tiers

5. Discussion

The integration of a tiered escalation framework with structured knowledge sharing has demonstrated substantial improvements in operational efficiency within cloud support environments. Quantitative findings underscore these gains, notably the 51% reduction in Mean Time to Resolution (MTTR) and a 23% increase in First Contact Resolution (FCR). These metrics reflect that the enhanced workflow effectively accelerates incident handling and empowers Tier 1 agents to address a broader range of issues without the need for escalation. The lowered escalation frequency further indicates that the framework not only resolves problems more quickly but also optimizes the utilization of skilled personnel across support tiers. By embedding knowledge-sharing mechanisms, the process enables frontline engineers to access contextualized solutions, troubleshoot with greater accuracy, and avoid repetitive or redundant escalations.

Beyond numerical improvements, qualitative insights reveal profound cultural and procedural impacts. Support engineers reported greater confidence, collaboration, and engagement after the framework's implementation. Tier 1 agents, previously dependent on higher-level engineers for complex issues, benefited from the structured access to shared knowledge and solution repositories. This not only improved their competence but also fostered a sense of ownership and professional growth. Simultaneously, Tier 2 and Tier 3 teams experienced relief from repetitive low-complexity tasks, allowing them to focus on high-priority and critical incidents that demand advanced expertise. The resulting distribution of workload aligns with best practices in operational excellence maximizing efficiency while ensuring each escalation tier contributes value where it is most needed.

The knowledge repository emerged as a pivotal component of this transformation. Serving as both a reference and a continuous learning platform, it ensured that every resolved issue contributed to organizational learning. Solutions captured in this repository enhanced resolution consistency and reduced knowledge silos, addressing a common challenge in large-scale support operations. The feedback loop also facilitated iterative improvement—engineers at all levels could refine solutions based on real-world applications, thereby sustaining the relevance and accuracy of shared information. Over time, this created a self-reinforcing ecosystem of expertise, collaboration, and innovation.

From an operational perspective, the framework presents scalable implications for enterprise-level cloud support teams. When augmented with AI-driven analytics, automated ticket triage, and collaboration tools, the model can further enhance productivity and service reliability. Integrating predictive analytics could enable proactive issue detection, minimizing downtime and enabling preemptive interventions. Moreover, embedding AI-based recommendation engines could assist support agents in real time, suggesting likely solutions or escalation paths based on historical data and context. Such extensions would align with the broader organizational shift toward intelligent operations and digital transformation.

However, the study acknowledges several limitations. The findings are specific to a single organization, and variations in culture, tools, and team structures across other cloud environments may yield different outcomes. Furthermore, the reliance on ticketing data and self-reported surveys introduces the possibility of bias or subjective interpretation. Despite these constraints, the study provides a strong foundation for further exploration.

6. Conclusion

The integration of tiered escalation with structured knowledge sharing has proven to be a highly effective strategy for enhancing cloud support efficiency. Quantitative improvements—such as a 51% reduction in Mean Time to Resolution (MTTR) and a 23% increase in First Contact Resolution (FCR) demonstrate measurable operational gains, while the decline in unnecessary escalations highlights the success of empowering frontline engineers through accessible knowledge resources. These outcomes collectively validate the framework's potential to streamline workflows, reduce resolution time, and elevate service quality in dynamic cloud environments.

Equally significant are the qualitative benefits observed within support teams. The framework cultivated a collaborative culture that encouraged continuous learning, confidence building, and stronger cross-tier communication. By converting individual expertise into a shared organizational asset, the knowledge repository not only improved incident handling consistency but also mitigated the recurring challenge of knowledge loss. The redistribution of workload across tiers enabled senior engineers to focus on strategic and high-complexity incidents, fostering overall operational balance and efficiency.

From a strategic standpoint, the framework offers scalability and adaptability across different organizational contexts. Integrating it with AI-based analytics, predictive monitoring, and automated categorization could further amplify its benefits, enabling proactive and intelligent support operations. Although the study's scope was limited to a specific organization, its findings provide a strong foundation for broader application and further research. In essence, the model demonstrates that a synergy between tiered escalation and knowledge sharing can transform cloud support operations into a more efficient, resilient, and learning-driven ecosystem.

Reference:

1. Addya, S. K., Satpathy, A., Chakraborty, S., & Ghosh, S. K. (2019). Optimal VM Coalition for Multi-Tier applications over Multi-Cloud Broker Environments. 2019 11th International Conference on Communication Systems & Networks (COMSNETS), 141–148. https://doi.org/10.1109/comsnets.2019.8711038
2. Baladari, V. (2022). Cloud resiliency engineering: Best practices for ensuring high availability in Multi-Cloud Architectures. International Journal of Science and Research (IJSR), 11(6), 2062–2067. https://doi.org/10.21275/sr220610115023

3. Chen, L., & Shen, H. (2016). Towards resource-efficient cloud systems: Avoiding over-provisioning in demand-prediction based resource provisioning. 2016 IEEE International Conference on Big Data (Big Data), 184–193. https://doi.org/10.1109/bigdata.2016.7840604
4. Fan, J., Wei, X., Wang, T., Lan, T., & Subramaniam, S. (2017). Deadline-aware task scheduling in a tiered IOT infrastructure. GLOBECOM 2017 - 2017 IEEE Global Communications Conference. https://doi.org/10.1109/glocom.2017.8255037
5. Fan, J., Wei, X., Wang, T., Lan, T., & Subramaniam, S. (2019). Churn-resilient task scheduling in a tiered IOT infrastructure. China Communications, 16(8), 162–175. https://doi.org/10.23919/jcc.2019.08.014
6. Hochleitner, M. (2020). The network of resources - facilitating a simplified and efficient exploitation of EO data in cloud environments. IOP Conference Series: Earth and Environmental Science, 509(1), 012024. https://doi.org/10.1088/1755-1315/509/1/012024
7. Khanna, R. (2016). IBM SmartCloud Cost Management with IBM cloud orchestrator cost management on the cloud. 2016 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM). https://doi.org/10.1109/ccem.2016.040
8. Mosayebi, M., & Azmi, R. (2023). Cost-Effective Clonal Selection and AIS-Based Load Balancing in Cloud Computing Environment. https://doi.org/10.21203/rs.3.rs-3077970/v1
9. Oh, K., Qin, N., Chandra, A., & Weissman, J. (2020). Wiera: Policy-driven multi-tiered geo-distributed cloud storage system. IEEE Transactions on Parallel and Distributed Systems, 31(2), 294–305. https://doi.org/10.1109/tpds.2019.2935727
10. Park, J., Wang, Q., Li, J., Lai, C.-A., Zhu, T., & Pu, C. (2016). Performance interference of memory thrashing in Virtualized Cloud Environments: A study of consolidated N-tier applications. 2016 IEEE 9th International Conference on Cloud Computing (CLOUD), 276–283. https://doi.org/10.1109/cloud.2016.0045
11. Paul, J. J. (2023). Serverless through the AWS well-architected framework. Distributed Serverless Architectures on AWS, 131–141. https://doi.org/10.1007/978-1-4842-9159-7_8
12. Raghavan, A., Chandra, A., & Weissman, J. B. (2014). Tiera. Proceedings of the 15th International Middleware Conference on - Middleware '14, 1–12. https://doi.org/10.1145/2663165.2663333
13. Rajasekaran, S., Duan, S., Zhang, W., & Wood, T. (2016). Multi-cache: Dynamic, efficient partitioning for multi-tier caches in Consolidated VM environments. 2016 IEEE International Conference on Cloud Engineering (IC2E), 182–191. https://doi.org/10.1109/ic2e.2016.10
14. Smirnov, A., Shilov, N., Ponomarev, A., & Schekotov, M. (2019). Human-computer cloud: Application platform and dynamic decision support. Proceedings of the 9th International Conference on Cloud Computing and Services Science, 120–131. https://doi.org/10.5220/0007725201200131
15. Suleiman, H., & Basir, O. (2019). Service level driven job scheduling in multi-tier cloud computing: A biologically inspired approach. 9th International Conference on Computer Science, Engineering and Applications (CCSEA 2019). https://doi.org/10.5121/csit.2019.90910