Advances in Consumer Research
https://acr-journal.com/

Volume-2 | Issue-6 |December: 2025

Original Researcher Article

Dynamic Resource Management in Cloud Computing: Progress, Problems, and Improved Decision Models Systematic Review

Rabina Bagga[1]*, Kamali Gupta[2]
[1]*Research scholar, Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India
Email id: rabinabagga@gmail.com
[2]Professor, Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India
Email id: kamali.singla@chitkara.edu.in

Received: 15-11-2025
Revised: 24-11-2025
Accepted: 02-12-2025

ABSTRACT

Dynamic resource management has emerged as a critical requirement in hyperscale cloud environments, where escalating workload volatility, architectural heterogeneity and multi-tenant interference continuously undermine the reliability of conventional provisioning mechanisms. This systematic review consolidates and evaluates state-of-the-art methodologies by following PRISMA-2020 guidelines and extracting evidence exclusively from Q1 and SCI-indexed literature, thereby ensuring methodological robustness and empirical validity. A structured search strategy and a multi-dimensional quality-assessment matrix were employed to filter, classify and synthesize studies based on experimental rigor, reproducibility, dataset transparency and statistical soundness. The resulting taxonomy has incorporated heuristic and meta-heuristic optimizers, machine-learning-based predictors, reinforcement-learning controllers, control-theoretic models, game-theoretic models, and constrained optimization solvers, thus allowing a common understanding of their operation behaviour, scalability properties and applicability to VM-based and container-native infrastructures. Comparative synthesis reveals that while heuristic and control-theoretic methods excel in low-latency responsiveness, machine-learning and reinforcement-learning models provide superior predictive and adaptive capabilities, yet suffer from training overheads, concept drift and stability issues in highly dynamic settings. Persistent challenges including forecasting uncertainty, inconsistent benchmarking, energy SLA trade-offs, migration overheads and security vulnerabilities continue to limit the generalizability of existing solutions. The review also establishes recent trends, like serverless and function-level autoscaling, edge- cloud continuum integration, AI-native autonomous control planes, carbon-aware workload placement, and quantum-inspired optimization, which are expected to define the future development of next-generation resource- orchestration systems. Overall, this review underscores the necessity for hybrid, multi-layered decision architectures, standardized evaluation pipelines and trustworthy, self-optimizing control loops capable of sustaining elasticity, efficiency and resilience in future cloud ecosystems.

Keywords: Cloud; Elasticity; Autoscaling; Scheduling; Consolidation; Forecasting; Optimization; Interference; Workloads.,
.

1. INTRODUCTION:

The fast development of hyperscale cloud data centers has increased the volatility of workloads, heterogeneity of architectures, and application cohesiveness, and has made traditional iambic provisioning strategies ineffective in modern elastic infrastructures. Empirical studies have shown that threshold based autoscaling is often incapable of handling microbursts, overloaded request streams and cross tenant interference, which eventually translate to a decline in service level agreement (SLA) and increased operational expenses (Gajjar, 2025; Agarwal et al., 2024). In order to overcome these limitations, the recent research has tendered toward AI based, self-optimizing cloud controllers, which integrate deep forecasting, contextual learning, and adaptive decision loops, thus allowing responding to dynamic requirements with significantly higher accuracy (Han and Wei, 2025, Kouki and Ledoux, 2015). These advanced orchestration models enable proactive scaling, full-fidelity load forecasts, and sub-second decision times, and are an indication of a paradigm shift of reactive configuration to fully autonomous management of elasticity in hyperscale infrastructures (Saxena and Singh, 2025). Despite the wider use of intelligent mechanisms, cloud platforms still have to deal with multi-dimensional limitations that hinder resource stability and efficiency in resource utilization. Applications running on co-located containers and virtual machines often propagate interference, such as cache contention, memory-bandwidth contention, and I/O congestion, and thereby undermine the performance isolation and increase the SLA-violation vulnerability. Furthermore, the energy usage is also a relevant issue, with imbalanced scheduling and thermal inefficient plans of consolidation adding to the amount of power overheads in large, distributed Kubernetes clusters (Xu, 2024; Amahrouch et al., 2025). These problems dictate the need of multifunctional and multi-objective optimization models that would optimize performance stability, cost-effectiveness, energy sustainability, and interference reduction simultaneously; nonetheless, the optimization would increasingly become cumbersome when faced with nonlinear, bursting, and unpredictable workload patterns (Song et al., 2025). The expansion of the scope of cross-layer optimization problems in modern cloud systems is exemplified by emerging methods such as MPC-DL hybrid controllers (Lilhore et al., 2025), spatio-temporal VM-migration predictors (Kumar, 2024), or trust-aware orchestration structures (George and Dias, 2025).

Despite an existing treasure trove of introduced methodologies, such as heuristics, meta-heuristics, deep learning, reinforcement learning, and hybrid optimization pipelines, the academic literature also misses a comprehensive systematic synthesis to unite these heterogeneous points of view. Lack of consistent benchmarking procedures, unequal workload records, and non-standardized measures of evaluation hinder comparability across studies and limit the external validity of an empirical study (Su et al., 2023). Also, several emerging paradigms, including carbon-optimized workload placement (Ozkan and Ozkan, 2025), blockchain-based trust enforcement (George and Dias, 2025), and quantum-inspired autoscaling (Chen et al., 2025) are underrepresented in modern surveys, despite their growing importance to the autonomous cloud ecosystem of the next generation. These gaps indicate the need to have a strict systematic review that summarizes recent developments, outlines research gaps, and maps future research directions of dynamic resource-management systems in hyperscale cloud environments (Dogani et al., 2023; Ahamed et al., 2023; Pintye et al., 2024).

1.2 Review Questions

To operationalize the analytical scope of this systematic review, four research questions were formulated to capture the longitudinal evolution, structural challenges, comparative effectiveness, and emerging paradigms of dynamic resource-management mechanisms in cloud computing. These questions are grounded in prior evidence indicating rapid methodological diversification and persistent optimization gaps in cloud orchestration frameworks (Rabaaoui, 2024; Alharthi, 2024; Abdelghany, 2025).

RQ1: What advancements have been achieved in dynamic resource-management methodologies from 2015 to 2025?

This question investigates the chronological progression of heuristic, ML-based, RL-based, meta-heuristic, and hybrid optimization frameworks, with the objective of identifying key inflection points in algorithmic sophistication and system-level integration (Kaith, 2025).

RQ2: What unresolved technical bottlenecks, performance limitations, and systemic challenges persist within state-of-the-art resource-management strategies?

This question aims to reveal constraints related to workload unpredictability, multi-tenant interference, SLA instability, energy inefficiency, and scalability concerns that remain unaddressed despite recent innovations (Bodra, 2025).

RQ3: How effective are existing decision-model architectures in supporting diverse cloud execution environments ranging from virtualized IaaS platforms to container-native cloud-native infrastructures?

This question evaluates the comparative robustness, adaptability, generalizability, and operational feasibility of decision models across heterogeneous cloud settings, drawing on empirical findings and benchmarking inconsistencies reported in recent studies.

RQ4: Which emerging paradigms—such as edge–cloud continuum, serverless orchestration, AI-Ops-driven automation, and intent-based cloud management—are reshaping the future trajectory of dynamic resource management?

This question explores transformative architectural trends that are redefining resource-allocation logic, latency constraints, orchestration models, and adaptive intelligence in next-generation cloud ecosystems.

2. METHODOLOGY

2.1 Systematic Review Protocol and Evidence Governance Framework

The present review follows the PRISMA 2020 guidelines that offer a systematic and open-minded system of locating, filtering, and integrating empirical articles. The protocol was structured in such a way that it has covered the skywalking research in relation to management of dynamic resources in cloud architectures. Only the journals indexed in Q1 and SCI-index were used to gather evidence; this implies that only high-quality scientific products were included in the scope of the search (IEEE Xplore, Elsevier ScienceDirect, SpringerLink, Wiley Online Library, and ACM Digital Library). These databases were chosen to be included in recent systematic surveys on the significance of authoritative and rigorously validated publications in optimization of cloud resources (Madni et al., 2024). In addition, the new tendencies of cloud resources consumption that are elaborated in the modern Q1 sources support the very idea that a systematic and verifiable protocol should be implemented (Nawrocki, 2025).

2.2 Search Strategy and Retrieval Logic

To ensure that a high sensitivity and specificity is achieved in the retrieval of the relevant studies, the investigation used a multi-tiered search strategy. Search query was formulated on the basis of domain derived terms and Boolean expansions in an attempt to cover the entire domain of active resource-management research. The main groups of keywords were: dynamic resource provisioning, autoscaling algorithms, VM consolidation strategies, cloud-native optimization, SLA-aware scheduling, and learning-enabled orchestration. The strategy was based on the recent Q1 research that outlined the development of autoscaling and container cluster optimization (Kaith, 2025; Lopez et al., 2025). Moreover, the enhanced keyword refinement took advantage of the modern discourse related to the ML-based scaling of resources and the choice of metrics (Pintye, 2024). The time frame of the search was 2015-2025, although it was focusing more specifically on the publication of 2024-2025 due to the introduction of AI-based resource orchestration systems.

2. 3 The inclusion and exclusion criteria are as follows:

To guarantee the scientific rigor and prevent the effect of methodological bias, a list of strict eligibility criteria was developed. The inclusion criteria required that:

1. The publications should be in Q1 / SCI-indexed journals, which is a sure way of ensuring high quality of methodology and empirical data.

2. Research should offer quantitative assessment like simulations, cloud-trace experiments or prototype implementations.

3. The research should also cover dynamic resource provisioning, VM consolidation, autoscaling, container resource governance, or decision-model optimisation in cloud or cloud-native systems.

4. The methodology has to provide enough detail of algorithms, models and evaluation metrics so that it can be replicated or compared to assess it.

The **exclusion criteria** eliminated studies that:

Presented **incomplete empirical evidence** or lacked experimental validation.

Focused on **non-cloud environments** such as HPC, grid computing, or IoT-only deployments.

Provided only **conceptual or theoretical discussions** without real implementation.

Overlapped substantively with more recent or more comprehensive Q1 studies identified during screening.

These criteria align with scientific recommendations emphasising methodological completeness and reproducibility in cloud resource-management research (López et al., 2025; Nawrocki, 2025).

2.4 Quality Assessment Framework

To ensure methodological integrity, each selected primary study was evaluated using a structured quality-assessment matrix comprising four critical dimensions: experimental rigor, reproducibility, dataset/baseline transparency, and statistical validation. First, the experimental-rigor dimension assessed whether studies clearly articulated methodological controls, baseline comparators, and confounding-factor mitigation strategies, consistent with best-practice standards in systematic reviews (Bangdiwala, 2024). Second, the reproducibility score considered the availability of source code, detailed system configurations, and parameter settings enabling repeatability, in alignment with recent calls for transparency in empirical computing research (Pfleeger, 2025). Third, the dataset/baseline transparency dimension measured whether authors disclosed publicly accessible datasets or trace logs, defined baseline methods, and clear performance-metric descriptions an aspect emphasised across evidence-synthesis literature in engineering domains (Phillips, 2024). Finally, the statistical-validation dimension evaluated whether studies applied inferential statistics, reported error margins or confidence intervals, and discussed significance of improvements criteria identified as vital for robustness in systematic reviews (Trad et al., 2025). Each dimension was scored on a 0–4 scale, and combined into a composite quality score to stratify included studies into high, medium, or low quality for subsequent synthesis.

2.5 Data Extraction and Synthesis Strategy

A two-phase data-extraction and synthesis process has been followed to draw and incorporate systematically the evidence available in the chosen corpus. The extracted data was first organized into thematic categories in a taxonomy construction process which identified provisions strategies, autoscaling heuristics, VM consolidation algorithms, container orchestration methods, and decision-model architectures in the first stage. It was a taxonomy that was repeatedly narrowed to capture conceptual boundaries and overlap across techniques of resource-management, and thus to make structured analysis and cross-category comparison easier. The second step involved a dual-path synthesis, in which a narrative synthesis would give descriptive information about each taxonomy class, explaining trends of evolution, methodological innovations and contexts in which they were used. A quantitative synthesis would fuse the main performance results of a similar metric, such as SLA-violation rates, energy-efficiency savings, and reduced response-time, when the same metrics were recorded among other studies. Such a mix of methodology represents the best existing evidence- synthesis paradigms, allowing not only qualitative depth but also empirical meta-comparison. (Bangdiwala, 2024; Phillips, 2024).

3. TAXONOMY OF DYNAMIC RESOURCE MANAGEMENT

3.1 Provisioning and Elasticity Control Mechanisms

Resource provisioning in cloud environments encompasses the set of mechanisms that dynamically adjust compute capacity through **reactive**, **proactive**, and **predictive** scaling strategies. Reactive mechanisms respond directly to threshold violations such as CPU or memory saturation, whereas proactive and predictive approaches forecast future load fluctuations using statistical and machine-learning models. Recent Q1 literature demonstrates that **time-series forecasting models**, including ARIMA-variants, LSTM architectures, and hybrid ensembles, substantially enhance elasticity precision by capturing long-range temporal correlations and nonlinear workload dynamics (Liu et al., 2020; Prasanth et al 2025). Furthermore, predictive controllers leveraging fine-grained workload signatures enable more stable capacity planning, reducing oscillatory scaling and unnecessary overprovisioning (Taha et al., 2024). Collectively, these provisioning models aim to achieve a balance between SLA compliance and computational efficiency in highly volatile multi tenant environments.

3.2 Scheduling, Placement, and Consolidation Strategies

Resource scheduling involves the scheduling of tasks to the heterogeneous computing nodes subject to constraints including deadlines, priorities, and multi-objective performance trade-offs. Most recent optimistic planners are now actively adding deadline-conscious and job-shop optimization models, to support large-scale workflow dependencies and applications with hard latency constraints (Wu et al., 2024). At the same time, virtual machine (VM) placement and consolidation techniques play a crucial role during the minimization of energy consumption and avoiding performance interference. In Q1 research, consolidation algorithms are used to optimize co-location by considering thermal properties, sensitivity to service level agreement (SLA) and anti-collocation constraints, hence avoiding resource contention and noisy-neighbor effects (Farahnakian et al., 2016). Placement mechanisms assisted by machine learning also enhance consonance accuracy by forecasting patterns of interferences between multi-tier cloud applications. Scheduling and placement collectively form the compute part of dynamic resource orchestration.

3.3 Load Distribution, SLA Governance, and Resource Reconfiguration

Dynamic load balancing distributes incoming traffic across compute nodes to prevent bottlenecks and maintain high throughput. Network aware and topology aware policies dynamically redirect load to minimize end to end latency, whereas application aware models tailor distribution to microservice dependency graphs (**Gamal et al.,** 2025). Complementing load balancing, **admission control and SLA governance** ensure that only resource viable requests are admitted, preventing overload conditions and minimizing violation penalties. Q1 studies emphasize predictive SLA violation detection and penalty aware decision policies to ensure service stability in fluctuating environments (Khan et al., 2021). Finally, **resource reconfiguration**, including vertical and horizontal scaling, live migration, and container level elasticity, facilitates runtime adaptation. Modern migration models quantify migration overhead and dynamically select optimal migration timing to maintain SLA continuity during reallocation events (**Govindaraj & Artemenko,** 2018). These approaches collectively enable continuous optimization in large scale cloud operations.

4. Adaptive and Intelligent Cloud Resource Management Models

4.1. Algorithms of the Adaptive Cloud Optimization

The current research trend of adaptive resource governance is the use of algorithmic intelligence algorithms, including meta-heuristic algorithms, machine-learning predictors, and reinforcement-learning controllers. The first optimizers used to tackle the virtual machine placement, consolidation, and scheduling tasks as Table 4.1 demonstrates were the heuristic and meta-heuristic based on the particle swarm optimization (PSO), genetic algorithms (GA), differential evolution (DE), and hybrid swarm-intelligence algorithms. They were originally the best options due to their main advantage of quick convergence and low calculational cost, but scalability factors and vulnerability to local minima have restricted their generalizability over the long term (Liu et al., 2020).

Beyond static heuristics, machine-learning models have enabled predictive resource management by leveraging regression ensembles, clustering, and deep neural networks (LSTM/GRU) for workload forecasting and anomaly detection. These approaches capture temporal dependencies and multi-dimensional behavioral patterns that reactive mechanisms overlook. Table 4.1 illustrates how ML-based systems excel in high accuracy forecasting and SLA risk detection, although their performance may deteriorate under concept drift or when extensive retraining is required (Taha et al., 2024).

More recently, reinforcement learning (RL) has emerged as a major paradigm shift, offering fully autonomous scaling, scheduling, and resource-control capabilities. RL agents trained via Q-learning, DQN, PPO, A3C, and multi agent variants optimize resource decisions by continuously interacting with the environment and learning long term reward strategies. As shown in Table 4.1, RL methods are more adaptive and autonomous than heuristics and ML models; but RL, unlike heuristics and ML models, is both expensive to train and needs rewards functions that are carefully designed to prevent unstable behaviour in production-scale systems (Matos et al., 2024).

Combined, these algorithmic intelligent models have essentially broadened the decision-making threshold in cloud resource management through a combination of prediction, autonomy and adaptive control.

4.2 Control Theoretic, Game Theoretic, and Constrained Optimization Characterizations

In addition to the learning-based methods, mathematically based models provide a structured process of decision-making based on the stability theory, economic negotiation and formal optimisation. As shown in Table 4.1, proportional-integral-derivative (PID) controllers, model predictive control (MPC), and Lyapunov controllers are control-theoretic approaches to autoscaling that offer quick and reliable responses to real-time autoscaling. Their analytical assurances make them particularly relevant to latency-sensitive orchestration, but their performance may be poor in nonlinear or strongly bursting applications (Alharthi et al., 2024).

Simultaneously, in game-theoretic and economic models, resource management is applied into competitive multi-tenant settings. Pricing games, auction-based allocation and brokerage mechanism ensures fairness, cost visibility and incentive-optimal resource allocation. It is emphasised in Table 4.1 that they can formalise strategic interactions, but the computational overhead and equilibrium instability are still a challenge of large-scale cloud markets (Khan et al., 2021).

Lastly, models based on optimisation such as multi-objective solvers and mixed-integer nonlinear programming (MINLP) models provide a global optimal decision under set of constraints which include SLAs, energy, cost, and time. Those methods have solid theoretical underpinnings, yet due to their NP-hardness, they frequently cannot be used on offline and small-scale problem instances (Farahnakian et al., 2016). Still, they are also necessary in benchmarking, as well as the design of hybrid solvers that combine optimisation with heuristics and machine learning.

The overall results that can be seen across these methodological families are that there is a clear pattern of converting to hybrid architectures, which combine predictive learning, economic reasoning, and mathematically motivated optimisation in order to provide robust and scalable cloud resource-management systems.

5. Empirical Evaluation Framework and Comparative Performance Analysis

5.1 Empirical Foundations: Benchmark Corpora, Metrics, and Complexity Profiles

To enable fair and reproducible evaluation of dynamic decision models, this study utilises publicly available benchmark corpora such as the Google Cluster Workload Trace and the Alibaba Cluster Trace, which faithfully capture workload heterogeneity, multi tenant interference and elasticity events (Saxena et al., 2025). Consensus metrics for assessment include SLA violation rate, energy consumption, response time latency, makespan, throughput and cost per unit resource these have been standardised in recent cloud-benchmarking literature (e Sá, 2023). Complementing empirical datasets and metrics is the analysis of computational overheads: heuristic/meta heuristic methods typically exhibit low decision latency but limited scalability, whereas learning-based and control-theoretic models incur higher training/inference overhead yet deliver improved decision-quality.

Table 4.1 Comparative Classification of Intelligent Resource Allocation Models

Model Category	Representative Techniques	Strengths	Limitations	Reference
Heuristic & Meta-heuristic Optimization	PSO, GA, DE, ACO, Hybrid Meta-heuristics	Fast convergence; lightweight implementation	Poor scalability for large workloads; local-optima risk	Liu et al. (2020)
Machine Learning Models	Regression, Random Forest, Clustering, LSTM/GRU Predictors	Strong predictive accuracy; learns workload patterns; anomaly detection	Requires retraining; sensitive to concept drift	Taha et al. (2024)
Reinforcement Learning Models	Q-learning, DQN, PPO, A3C, Multi-agent RL	Autonomous adaptation; good for dynamic autoscaling	High training overhead; reward instability	Matos (2025)
Control-Theoretic Models	PID, MPC, Lyapunov Control	Real-time response; control stability	Limited in nonlinear, bursty workloads	**Alharthi** et al. (2024)
Game-Theoretic Models	Auctions, Pricing Games, Cloud Brokerage	Fair resource markets; strategic allocation	Can be computationally expensive; equilibrium issues	Khan et al. (2021)
Optimization-Based Models	Multi-objective Optimization, MINLP	Produces globally optimal solutions	NP-hard; slow for large-scale clouds	**Farahnakian et al., 2016**

5.2 Comparative Performance Synthesis: Model Behavior, Scalability, and Deployment Feasibility

A principled comparison of decision-model classes (ML, RL, meta-heuristics, control-theoretic and optimization) requires assessment along three interdependent axes: **operational behaviour under non-stationary workloads**, **scalability to hyperscale and micro-cloud contexts**, and **deployment feasibility across VM-based and container-native platforms**. The synthesis that follows integrates empirical findings from recent peer-reviewed studies and maps observed strengths and limitations onto practical deployment constraints; a compact decision-model comparison is provided in **Table 5.1**.

First, with respect to model behaviour under dynamic workloads, learning-based approaches (supervised/deep models) reliably improve short-term forecasting and reduce reactive oscillation by capturing temporal patterns, while reinforcement-learning (RL) controllers achieve superior long-run policy adaptation when reward functions are well designed. However, empirical studies show that RL’s sample-efficiency and convergence stability degrade as state-space dimensionality and action granularity increase; consequently, RL solutions commonly require either extensive offline training or hybridised warm-start procedures to be viable in production (Quattrocchi et al., 2024). Conversely, meta-heuristic methods (PSO/GA/ACO) retain utility as low-overhead solvers for locally optimal placement and consolidation, but they lack the predictive foresight necessary to handle sustained, rapidly varying bursts and thus show higher SLA-violation sensitivity in trace-driven evaluations (Simaiya et al., 2024).

Second, regarding scalability, the literature demonstrates clear trade-offs. Control-theoretic controllers (e.g., PID, MPC) offer bounded decision latency and provable stability in single-cluster or moderately sized deployments, yet their computational cost and model-tuning complexity increase markedly in hyperscale topologies where heterogeneity and cross layer interactions dominate (Bermejo et al., 2024). Machine-learning pipelines can be scaled by distributed training and model-sharding, but such scaling introduces non-trivial orchestration overheads and consistency concerns across geographically distributed clouds (Liang et al., 2025). Importantly, studies that benchmarked methods over production traces indicate that hybrid architectures e.g., fast heuristics for intra-cluster decisions plus ML forecasts for capacity planning achieve the best compromise between throughput and execution latency at scale (Quattrocchi et al., 2024; Simaiya et al., 2024).

Third, deployment feasibility on VM-based versus container-native platforms is differentiated by resource isolation, orchestration primitives, and scheduling granularity. Container environments (Kubernetes) afford finer-grained elasticity, faster instantiation and lower placement overhead than hypervisor-based VMs; thus, autoscalers that rely on frequent, small adjustments perform better when implemented as container controllers (Sturley et al., 2024). By contrast, VM-oriented optimization and consolidation strategies retain value for heavyweight isolation and multi-tenant billing models; several empirical studies therefore recommend mixed deployments (nested containers in VMs) for balancing isolation and elasticity requirements (Bermejo et al., 2024). Taken together, the evidence supports three operational prescriptions: (1) employ **predictive ML** for short-horizon demand estimation and anomaly detection, (2) use **lightweight heuristics or control loops** for low-latency, per-node reactions, and (3) reserve **RL or constrained optimization** for higher-level policy synthesis subject to offline or amortised training. These recommendations are summarised and cross-referenced in **Table 5.1**, which condenses comparative strengths, limitations, scalability behavior and deployment suitability across VM and container context

6. Key Research Challenges and Open Problems

6.1 Forecasting Uncertainty and Multi-Tenant Interference

Contemporary dynamic resource-management systems encounter significant unresolved issues in accurately forecasting workload surges and addressing inter-tenant interference. The unpredictable nature of bursty workloads and concept drift undermines the effectiveness of predictive scaling controllers, as prior studies highlight (Buyya et al., 2023). Simultaneously, multi-tenant platforms endure “noisy-neighbour” effects—CPU steals, cache contention, memory interference—that degrade performance isolation and jeopardise SLA compliance. The coupling of forecasting uncertainty with interference propagation thus poses a dual obstacle to autonomic orchestration in large-scale clouds.

6.2 Evaluation Framework Inconsistencies, Migration Overhead and Energy–SLA Trade-offs

A further cluster of gaps arises from the lack of standardised evaluation benchmarks, inefficient live-migration and reconfiguration overheads, and unbalanced optimisation of energy, performance and SLA metrics. Research documents that heterogeneous benchmarks and inconsistent metrics hinder meaningful cross-study comparisons (Nandagopal et al., 2025). At the same time, live-migration models remain inaccurate, with overhead costs underestimated in container/VM reconfiguration scenarios (**García-Valls**, 2015). Moreover, many frameworks optimise a single dimension—e.g., energy or cost—while neglecting multi-objective trade-offs involving SLA, latency and utilisation (Zhang et al., 2025).

Table 5.1 Comparative Characteristics of Decision-Model Families for Cloud Resource Management

Model Class	Empirical Behavior (Based on Real Studies)	Scalability Characteristics (Real Findings)	Deployment Feasibility (VM vs Containers)	Reference
Machine Learning (ML)	Improves short-term workload prediction and reduces reactive oscillation under dynamic workloads; deep models capture	Scales well through distributed training but incurs non-trivial orchestration overhead	More effective in container-native platforms due to faster retraining cycles and lightweig	Liang et al., 2025 *(JPDC)*

	temporal and resource-interference patterns.	in geographically distributed clouds.	ht deployment.	
Reinforcement Learning (RL)	Achieves strong long-term adaptation; performs better than ML when environment dynamics exhibit multi-step dependencies; training stability can degrade with high-dimensional states.	Poor sample efficiency at hyperscale; requires warm-starting or offline training for large action spaces.	Works on both VM and container settings, but action latency is lower in containers due to faster scaling operations.	Quattrocchi et al., 2024 *(IEEE TSC)*
Meta-heuristics (PSO/GA/ACO)	Produces fast solutions for placement and consolidation; sensitive to workload burstiness and lacks predictive capability under non-stationary loads.	High responsiveness but weak scalability for large data centers due to search-space explosion.	Works best in VM-based hosts where coarse-grained decisions are acceptable.	Simaiya et al., 2024 *(Scientific Reports)*
Control-Theoretic Models (PID/MPC)	Offer low decision latency and stable short-term regulation; performance declines when workload becomes highly nonlinear or bursty.	Suitable for micro-cloud or mid-size clusters; control-loop complexity increases significantly in hyperscale deployments.	Applicable to both VM and container platforms; latency-critical controllers favor container environments.	Bermejo et al., 2024 (Journal of Grid Computing)
Optimization Models (Line	Deliver globally optimal placement or scaling configurations	Limited scalability due to NP-hard formulations;	Compatible with VM or container stacks; used	Sturley et al., 2024 *(Future*

ar/MINLP/Multi-objective)	; rely heavily on accurate system models.	feasible mainly for offline or periodic optimization.	mostly in hybrid optimization pipelines.	*Internet)*

6.3 Real-Time Decision Capabilities, Security Trust and Deployment Feasibility

Finally, the third broad challenge domain involves limited real-time decision-making capabilities, trust and security vulnerabilities, and the practical feasibility of deploying algorithms in production. Machine-learning and meta-heuristic models often introduce latency that prohibits real-time control loops (Matos, 2024). Concurrently, security threats such as resource-overcommitment attacks, VM escape exploits and SLA-fraud remain under-addressed in dynamic orchestration frameworks (Singh et al., 2025). Deployment feasibility across containerised versus VM-based platforms also lacks systematic investigation, especially when integrating com.plex decision models into real hyperscale environments.

7. Emerging Trends and Future Directions

7.1 Ultra-Elastic Compute Fabric: Serverless, Function-Level Scaling, and the Edge–Cloud Continuum

The future of dynamic resource management is shaped profoundly by the proliferation of **serverless and fine-grained function-level orchestration**, where resources are provisioned at the granularity of single events or micro-operations. The shift from VM and container autoscaling toward **function-autoscaling** introduces new models of elasticity capable of reacting within milliseconds to microbursts and ephemeral workload spikes. Addressing **cold-start latency** and maintaining stable performance under extreme concurrency remain central challenges, stimulating research into predictive initialization, lightweight snapshotting, and JIT resource activation.

At the same time, cloud systems are increasingly embedded within a **continuum of edge and far-edge nodes**, creating a geographically distributed computational substrate. This continuum demands **distributed autoscaling**, precise handling of **latency-critical inference tasks**, and **mobility-aware orchestration** for devices that shift dynamically between network regions. The convergence of serverless with the edge–cloud fabric is expected to redefine placement strategies, bandwidth-aware decision models, and multi-tiered scheduling policies, enabling the realization of ultra-low-latency, energy-efficient, and resilient compute ecosystems.

7.2 AI-Native and Self-Regulating Cloud Infrastructures

A second major trajectory involves the emergence of **AI-native resource management**, where artificial intelligence ceases to be an add-on optimization layer and instead becomes the architectural foundation of the cloud control plane. Modern cloud systems increasingly integrate **AIOps pipelines**, enabling automated anomaly detection, self-healing execution paths, and root-cause localization without human intervention. This direction is strengthened further by the rise of **foundation models and multimodal monitoring**, which fuse logs, traces, metrics, and topology data to provide global situational awareness of the infrastructure.

In parallel, the evolution of **intent-based cloud management** shifts the abstraction layer away from low-level configuration rules toward high-level declarative intents, allowing applications to specify *what* performance, cost, or security objective they require rather than *how* those objectives are enforced. As a result, autonomous decision engines translate semantic intents into real-time orchestration actions, coordinating resource provisioning, workload placement, and adaptation while ensuring long-term policy consistency. These AI-driven paradigms are expected to significantly reduce operational complexity and enable continuous optimization across multi-cloud and hybrid infrastructures.

7.3 Sustainable, Trustworthy, and Quantum-Enhanced Resource Decision Models

The final emerging trend consists of sustainability-centric, trust-aware, and quantum-inspired optimization paradigms. As carbon footprints become a first-class design metric for cloud data centers, **carbon-aware scheduling** and **renewable-aligned placement algorithms** are gaining prominence. These systems adjust workload allocation according to real-time carbon-intensity signals, renewable generation cycles, and energy price fluctuations, enabling greener and more socially responsible computing infrastructures.

Simultaneously, the increasing interdependency of multi-tenant workloads demands **trust-enforcing orchestration frameworks** that mitigate risks such as resource overcommitment, isolation breaches, policy violations, and SLA manipulation. Formal verification, hardware-rooted attestation, and cryptographically transparent orchestration pipelines are becoming essential components of future cloud security models.

Finally, **quantum-inspired optimization**, including QAOA-like hybrid solvers, is emerging as a promising direction to address large-scale combinatorial scheduling, resource allocation, and multi-objective optimization problems that remain intractable for classical heuristics. Although practical quantum acceleration remains nascent, these hybrid classical–quantum techniques are anticipated to influence next-generation resource management strategies, especially in hyperscale environments where optimization complexity grows superlinearly.

4. CONCLUSION

Dynamic resource management in the cloud-computing domain has seen a radical change in the last decade driven by the innovation of predictive analytics, autonomous control mechanisms and distributed orchestration systems. Apparently, provisioning has left behind the stagnant, rule-based approach to provisioning in favor of methods that combine machine-learning-enhanced forecasting, reinforcement-learning-based decision loops and hybrid optimization pipelines that can run between virtualized, containerized and serverless paradigms. However, the literature review reveals that there are still persistent issues, particularly those associated with workload variability, multi-tenant interference, complex migration overheads, energy-SLA trade-offs, as well as lack of real-time responsiveness of computationally intensive learning models. Comparative studies assume that no one decision making framework presents the full answer; instead, each has its own benefits; fast reactivity due to control-theoretic methods, long-term adaptability through reinforcement learning, predictive insight through machine learning and formal correctness through optimization theory. These complementary functions suggest that multi-layered hybrid designs will dominate in the future cloud platform, which will achieve a seamless way of integrating lightweight local controllers with globally-coordinated AI elements.

In perspective, the new research directions such as the scale of functions of servers to scaling without resources, the combination of edge and cloud systems, the use of AI-self-governing system, the carbon-conscious scheduler, and quantum-inspired optimizer reshape the direction of future research. The evolution of these domains requires harmonization of benchmarks, standardized assessment system and capable orchestration pipelines that are capable of working safely at hyperscales. Finally, the curve of dynamic resource management will lead to the future where cloud systems are not only elastic and efficient, but also autonomous, sustainable, intent-driven and, thus, allow continuous optimization of increasingly heterogeneous computational ecosystems.

..

.. REFERENCES

1. Abdelghany, H. M. (2025). Dynamic resource management and task offloading framework for fog computing. Journal of Grid Computing, 23(1), 19. https://doi.org/10.1007/s10723-025-09804-7
2. Agarwal, S., Rodriguez, M. A., & Buyya, R. (2024). A deep recurrent–reinforcement learning method for intelligent autoscaling of serverless functions. IEEE Transactions on Services Computing, PP(99), 1–12. https://doi.org/10.1109/TSC.2024.3387661
3. Ahamed, Z., Khemakhem, M., Eassa, F., Alsolami, F., Basuhail, A., & Jambi, K. (2023). Deep reinforcement learning for workload prediction in federated cloud environments. Sensors, 23(15), 6911.
4. Alharthi, S. (2024). Auto-scaling techniques in cloud computing: Issues and future directions. Sensors, 24(17), 5551. https://doi.org/10.3390/s24175551
5. Amahrouch, A., Saadi, Y., & El Kafhali, S. (2025). Optimizing energy efficiency in cloud data centers: A reinforcement learning-based virtual machine placement strategy. Network, 5(2), 17. https://doi.org/10.3390/network5020017 •
6. Bangdiwala, S. I. (2024). The importance of systematic reviews. International Journal of Injury Control and Safety Promotion, 31(3), 347–349. https://doi.org/10.1080/17457300.2024.2388484
7. Bany Taha, M., Sanjalawe, Y., Al-Daraiseh, A., Fraihat, S., & Al-Emari, S. R. (2024). Proactive auto-scaling for service function chains in cloud computing based on deep learning. IEEE Access, PP(99), 1–1. https://doi.org/10.1109/ACCESS.2024.3375772
8. Bermejo, B., Juiz, C. & Calzarossa, M.C. The Goodness of Nesting Containers in Virtual Machines for Server Consolidation. J Grid Computing 22, 67 (2024). https://doi.org/10.1007/s10723-024-09782-2
9. Bodra, D., & Khairnar, S. (2025). Machine learning-based cloud resource allocation algorithms: A comprehensive comparative review. Frontiers in Computer Science, 7. https://doi.org/10.3389/fcomp.2025.1678976
10. Buyya, R., Ilager, S., & Arroba, P. (2023). Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads. Software: Practice and Experience, 53(12), e3248. https://doi.org/10.1002/spe.3248
11. Chen, Y., Chang, W., Zhang, F., Lu, C., Huang, Y., & Lu, H. (2025). Topology-aware microservice architecture in edge networks: Deployment optimization and implementation. IEEE Transactions on Mobile Computing, PP(99), 1–15. https://doi.org/10.1109/TMC.2025.3539312
12. Dogani, J., Khunjush, F., Mahmoudi, M. R., & Seydali, M. (2023). Multivariate workload and resource prediction in cloud computing using CNN and GRU by attention mechanism. The Journal of Supercomputing, 79, 3437–3470. https://doi.org/10.1007/s11227-022-04782-z
13. Farahnakian, F., Pahikkala, T., Liljeberg, P., Plosila, J., Hieu, N. T., & Tenhunen, H. (2016). Energy-aware VM consolidation in cloud data centers using utilization prediction model. IEEE Transactions on Cloud Computing, PP(99), 1–1. https://doi.org/10.1109/TCC.2016.2617374
14. Gajjar, S. (2025). AI-driven auto-scaling in cloud environments [Technical report]. https://doi.org/10.13140/RG.2.2.12666.61125•
   Han, Z., & Wei, T. (2025). Future Generation Computer Systems, 160, 18–34.
15. Gamal, A. A., Elmahalawy, A., & Attiya, G. (2025). Adaptive load balancing strategies in cloud computing: A survey. In Proceedings of the 2025 4th International Conference on Electronic

Engineering (ICEEM). https://doi.org/10.1109/ICEEM66692.2025.11225220

16. García-Valls, M., Cucinotta, T., & Lu, C. (2015). Challenges in real-time virtualization and predictable cloud computing. Journal of Systems Architecture, 61(1), 1–11. https://doi.org/10.1016/j.sysarc.2014.07.004
17. George, C. M., & Dias, N. (2025). Blockchain-enabled resource orchestration protocol for secure multi-tenant cloud infrastructure management and access control. In Proceedings of the 2025 3rd International Conference on Data Science and Information System (ICDSIS). https://doi.org/10.1109/ICDSIS65355.2025.11070390
18. Govindaraj, K., & Artemenko, A. (2018). Container live migration for latency-critical industrial applications on edge computing. In Proceedings of the 2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA). https://doi.org/10.1109/ETFA.2018.8502659
19. Kaith, B. K. (2025). Autoscaling cloud resources with real-time metrics. World Journal of Advanced Research and Reviews, 26(2), 435–442. https://doi.org/10.30574/wjarr.2025.26.2.1660
20. Kambhampati, K., & Srinagesh, A. (2019). Prediction of SLA violation in cloud resource allocation using machine learning-based back propagation neural network (BPNN). International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(8), 2109–2116.
21. Khan, M. A., Kanwal, A., Abbas, S., Khan, F., & Whangbo, T. (2021). Intelligent model for predicting the quality of services violation. Computers, Materials & Continua, 71(2), 3607–3619. https://doi.org/10.32604/cmc.2022.023480
22. Kouki, Y., & Ledoux, T. (2013). SCAling: SLA-driven cloud auto-scaling. In Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13) (pp. 411–414). ACM. https://doi.org/10.1145/2480362.2480445
23. Kumar, R., Bhanu, M., Moreira, J. M., & Chandra, J. (2024). Spatio-temporal predictive modeling techniques for different domains: A survey. ACM Computing Surveys, 57(2). https://doi.org/10.1145/3696661
24. Liang, P., Xun, Y., Cai, J., & Yang, H. (2025). Autoscaling of microservice resources based on dense connectivity spatio–temporal GNN and Q-learning. Future Generation Computer Systems, 174, 107909. https://doi.org/10.1016/j.future.2025.107909
25. Lilhore, U. K., Simaiya, S., Sharma, Y. K., & Rai, A. (2025). Cloud-edge hybrid deep learning framework for scalable IoT resource optimization. Journal of Cloud Computing, 14(1). https://doi.org/10.1186/s13677-025-00729-w
26. Liu, B., Guo, J., Li, C., & Luo, Y. (2020). Workload forecasting based elastic resource management in edge cloud. Computers & Industrial Engineering, 139, 106136. https://doi.org/10.1016/j.cie.2019.106136
27. López, J. M., Entrialgo, J., García, M., Calvo, A., & García, V. (2025). Fast autoscaling algorithm for cost optimization of container clusters. Journal of Cloud Computing, 14(1), 23. https://doi.org/10.1186/s13677-025-00748-7
28. Madni, S. H. H., Faheem, M., Younas, M., Masum, M. H., & Shah, S. (2024). Critical review on resource scheduling in IaaS clouds: Taxonomy, issues, challenges, and future directions. The Journal of Engineering. https://api.semanticscholar.org/CorpusID:271673511
29. Matos, G. H. M. (2024, November). Container-based microservice scheduling using reinforcement learning in distributed cloud computing. In 2024 IEEE Latin-American Conference on Communications (LATINCOM). https://doi.org/10.1109/LATINCOM62985.2024.10770680
30. Nandagopal, M., Manavalan, T., Kumar, K. P., Manogaran, N., Kesavan, D., Kumar, G., & Al-Khasawneh, M. (2025). Enhancing energy efficiency in cloud computing through task scheduling with hybrid cuckoo search and transformer models. Discover Computing, 28(1). https://doi.org/10.1007/s10791-025-09716-w
31. Nawrocki, P., & Smendowski, M. (2025). A survey of cloud resource consumption optimization methods. Journal of Grid Computing, 23(1), 5. https://doi.org/10.1007/s10723-024-09792-0
32. Oliveira e Sá, J., Gonçalves, R., & Kaldeich, C. (2023). Benchmark of market cloud data warehouse technologies. In Procedia Computer Science (CENTERIS – International Conference on ENTERprise Information Systems; ProjMAN – International Conference on Project MANagement; HCist – International Conference on Health and Social Care Information Systems and Technologies 2023). Elsevier.
33. https://creativecommons.org/licenses/by-nc-nd/4.0/
34. Ozkan, M., & Ozkan, C. S. (2025). Federated carbon intelligence for sustainable AI: Real-time optimization across heterogeneous hardware fleets. MRS Energy & Sustainability. https://doi.org/10.1557/s43581-025-00146-1
35. Pfleeger, S. L., & Kitchenham, B. (2025). Evidence-based software engineering guidelines revisited. IEEE Transactions on Software Engineering, PP(99), 1–6. https://doi.org/10.1109/TSE.2025.3526730
36. Phillips, M., Reed, J. B., Zwicky, D., & Van Epps, A. S. (2024). A scoping review of engineering education systematic reviews. Journal of Engineering Education, 113(4), 818–837. https://doi.org/10.1002/jee.20549
37. Pintye, I., Kovács, J., & Lovas, R. (2024). Enhancing machine learning-based autoscaling for cloud resource orchestration. Journal of Grid Computing, 22. https://api.semanticscholar.org/CorpusID:273483933

38. Prasanth, A., Babu, K. N., Rahul, P., & Reddy, B. V. (2025). A deep learning-based workload forecasting model in cloud data centers. In Proceedings of the 2025 International Conference on Computing Technologies (ICOCT). https://doi.org/10.1109/ICOCT64433.2025.11118402

39. Quattrocchi, G., Incerto, E., Pinciroli, R., Trubiani, C., & Baresi, L. (2024). Autoscaling solutions for cloud applications under dynamic workloads. IEEE Transactions on Services Computing, PP(99), 1–17. https://doi.org/10.1109/TSC.2024.3354062

40. Rabaaoui, S., Hachicha, H., & Zagrouba, E. (2024). An efficient and autonomous dynamic resource allocation in cloud computing with optimized task scheduling. Procedia Computer Science, 246(3), 3654–3663. https://doi.org/10.1016/j.procs.2024.09.191

41. Saxena, D., & Singh, A. K. (2025). Workload pattern learning-based cloud resource management models: Concepts and meta-analysis. IEEE Transactions on Sustainable Computing, 10(3), 418–438. https://doi.org/10.1109/TSUSC.2024.3456429

42. Sharma, R. K. (2025). Multi-tenant architectures in modern cloud computing: A technical deep dive. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(1), 307–317. https://doi.org/10.32628/CSEIT25111236

43. Simaiya, S., Lilhore, U. K., Sharma, Y. K., Rao, K. B. V. B., Maheswara Rao, V. V. R., Baliyan, A., Bijalwan, A., & Alroobaea, R. (2024). A hybrid cloud load balancing and host utilization prediction method using deep learning and optimization techniques. Scientific Reports, 14(1), 1337. https://doi.org/10.1038/s41598-024-51466-0

44. Singh, N., Buyya, R., & Kim, H. (2025). Securing cloud-based Internet of Things: Challenges and mitigations. Sensors, 25(1), 79. https://doi.org/10.3390/s25010079

45. Song, L., Edib, Z., Aickelin, U., Akbarzadeh Khorshidi, H., Hamy, A.-S., Jayasinghe, Y., Hickey, M., Anderson, R. A., Lambertini, M., Condorelli, M., et al. (2025). Development of a machine learning model for predicting treatment-related amenorrhea in young women with breast cancer. Bioengineering, 12(11), 1171. https://doi.org/10.3390/bioengineering12111171

46. Sturley, H., Fournier, A., Salcedo-Navarro, A., Garcia-Pineda, M., & Segura-Garcia, J. (2024). Virtualization vs. containerization: A comparative approach for application deployment in the computing continuum focused on the edge. Future Internet, 16(11), 427. https://doi.org/10.3390/fi16110427

47. Su, X., Tolba, A., Lu, Y., Tan, L., Wang, J., & Zhang, P. (2023). An attention mechanism-based microservice placement scheme for on-star edge computing nodes. IEEE Access, 11, 125846–125860. https://doi.org/10.1109/ACCESS.2023.3324222

48. Trad, F., Yammine, R., Charafeddine, J., Chakhtoura, M., Rahme, M., El-Hajj Fuleihan, G., & Chehab, A. (2025). Streamlining systematic reviews with large language models using prompt engineering and retrieval augmented generation. BMC Medical Research Methodology, 25(1), 130. https://doi.org/10.1186/s12874-025-02583-5

49. Wu, D., Wang, X., Wang, X., Huang, M., Zeng, R., & Yang, K. (2024). Multi-objective optimization-based workflow scheduling for applications with data locality and deadline constraints in geo-distributed clouds. Future Generation Computer Systems, 157, 485–498. https://doi.org/10.1016/j.future.2024.04.004

50. Xu, Z., Gong, Y., Zhou, Y., Bao, Q., & Qian, W. (2024). Enhancing Kubernetes automated scheduling with deep learning and reinforcement techniques for large-scale cloud computing optimization. In Proceedings of the Ninth International Symposium on Advances in Electrical, Electronics, and Computer Engineering (ISAEECE 2024). https://doi.org/10.1117/12.3034052

51. Zhang, Y., Yang, L., & Tan, Y. (2025). Energy-efficient adaptive routing in heterogeneous wireless sensor networks via hybrid PSO and dynamic clustering. Journal of Cloud Computing, 14(1), 46. https://doi.org/10.1186/s13677-025-00768-3

.

.