Transactions on Machine Intelligence 9(1) (2026) 16–31

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI

A Systematic Review of Fairness-Aware and Optimization-Driven Virtual Machine Allocation Models in Cloud Computing

O. Johnson Tope[1,*]

[1] MSc ,Computer Science, Achievers University, Owo , Nigeria

ARTICLE INFO

Article History:
Received 9 October 2025
Received in revised form 28 December 2025
Accepted 24 February 2026
Available online 2 March 2026

Keywords:

Cloud computing, Virtual machine allocation, Fairness, Multi-objective optimization, Proportional equity, Resource management, Load balancing, Systematic review.

ABSTRACT

Virtual machine (VM) allocation remains a foundational challenge in Infrastructure-as-a-Service (IaaS) cloud computing, particularly as multi-tenant environments grow in scale, heterogeneity, and complexity. Recent advancements increasingly emphasize the dual objectives of fairness and optimization, reflecting the need for equitable resource distribution without compromising system performance. This systematic review synthesizes peer-reviewed research published between 2020 and 2025, evaluating fairness-driven, optimization-driven, and hybrid VM allocation models. Following a PRISMA-aligned search strategy across IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar, 15 high-quality studies meeting strict inclusion criteria were selected. Findings reveal that while fairness-driven models successfully reduce resource starvation and improve equity, they often incur higher makespan and computational overhead. Conversely, optimization-driven models demonstrate substantial gains in throughput, energy efficiency, and load balancing, yet frequently deprioritize fairness, leading to disproportionate resource allocation for low-priority tasks. Hybrid models have subsequently emerged as a dominant research direction, leveraging multi-objective and machine learning-based approaches to balance efficiency and equity effectively. Despite this progress, key research gaps persist, including a lack of standardized fairness metrics, limited validation using real-world workload traces, and inconsistent reporting of scalability and convergence properties. This review situates the Optimized Proportional Equity Model (OPEM) within this evolving landscape, highlighting its deterministic, fairness-centered design and potential to contribute to transparent resource allocation frameworks. The study concludes with actionable recommendations for developing next-generation allocation mechanisms grounded in rigorous evaluation, hybrid optimization, and standardized assessment.

1. INTRODUCTION

Cloud computing has firmly established itself as the premier paradigm for provisioning elastic, scalable, and on-demand computational services, with Virtual Machine (VM) allocation serving as a foundational operational mechanism within Infrastructure-as-a-Service (IaaS) platforms. The process of VM allocation dictates how

* Corresponding Author: johnsonolujayetan@yahoo.com
MSc ,Computer Science, Achievers University, Owo , Nigeria

http://dx.doi.org/10.22034/TMI.2026.244942

heterogeneous hardware capabilities specifically CPU cycles, volatile memory, physical storage, and network bandwidth are distributed among competing users and highly dynamic workloads in multi-tenant environments. As cloud workloads expand in scale, complexity, and variability, engineering efficient, fair, and cost-conscious allocation strategies has become vital. These strategies are essential for sustaining Quality of Service (QoS), minimizing provider operational overhead, and guaranteeing strict compliance with Service-Level Agreements (SLAs). Recent literature emphasizes that resource orchestration remains one of the primary determinants governing global cloud performance, data center energy efficiency, and end-user satisfaction [1-2].

Modern cloud infrastructures face escalating complexity driven by heterogeneous application workloads, non-linear demand spikes, and the geographical dispersion of massive, hyper-scale data centers. These conditions introduce severe technical challenges for traditional VM allocation pipelines, including severe resource contention, stochastic workload shifts, latency variability, and the systemic risk of resource starvation for lower-priority tasks. Conventional heuristic, greedy, or static threshold-based allocation frameworks frequently fail to adapt to such non-stationary, highly volatile cloud environments. Consequently, contemporary research trajectories have shifted toward optimization-driven methodologies. These encompass evolutionary meta-heuristics, mathematical programming, deep reinforcement learning, and hybrid intelligent systems, all designed to enhance the real-time adaptability and structural responsiveness of cloud schedulers [3-4].

Alongside raw performance optimization, structural fairness has emerged as a critical metric within shared, multi-tenant cloud ecosystems. Fairness-driven allocation paradigms such as proportional fairness, max-min fairness, and equity-oriented resource distribution are explicitly engineered to prevent resource monopolization by dominant tenants, thereby ensuring predictable, isolated service delivery across diverse user classes. Recent Scopus-indexed literature highlights the strategic importance of fairness in preventing user-level performance degradation and maintaining equitable resource availability under fluctuating network demands [5-6]. Nevertheless, implementing fairness-aware strategies frequently introduces complex multi-objective trade-offs, often elevating computational complexity, degrading global system throughput, or diminishing VM consolidation and energy efficiency.

Despite substantial methodological progress, several critical challenges remain unresolved in the literature. A significant limitation is that many existing VM allocation models treat fairness as a secondary, implicit objective, embedding it loosely within broader single-objective optimization frameworks rather than designing explicit, equity-oriented algorithmic mechanisms. Furthermore, the absence of standardized fairness metrics and the prevalence of inconsistent experimental environments complicate rigorous cross-model comparative evaluation, thereby hindering real-world industrial adoption. These limitations underscore the critical need for a systematic, comprehensive synthesis of the fairness-aware and optimization-driven VM allocation paradigms published in recent years.

This review systematically addresses these literature gaps by providing a critical taxonomy and evaluation of optimization-based, fairness-oriented, and hybrid VM allocation architectures developed between 2020 and 2025. It identifies emerging methodological innovations, core performance trade-offs, and unresolved empirical challenges, while positioning the Optimized Proportional Equity Model (OPEM) as a state-of-the-art framework that successfully harmonizes formal fairness guarantees with operational throughput efficiency. Through this systematic synthesis, this study provides actionable insights to support knowledge engineers in architecting transparent, equitable, and highly scalable VM allocation mechanisms tailored for next-generation cloud environments.

2. BACKGROUND AND CONCEPTUAL FRAMEWORK

2.1. Virtual Machine Allocation in Cloud Computing

Virtual machine (VM) allocation is a foundational mechanism in cloud computing and is responsible for mapping computational tasks, commonly referred to as cloudlets, to available virtualized resources. These resources include CPU cycles, memory, storage, and network bandwidth as shown in Figure 1, each of which must be distributed efficiently to satisfy users' service-level expectations and optimize overall system performance [7-8]. The inherent heterogeneity of cloud workloads, coupled with the dynamic nature of multi-tenant IaaS environments, demands allocation strategies that are both adaptive and scalable.

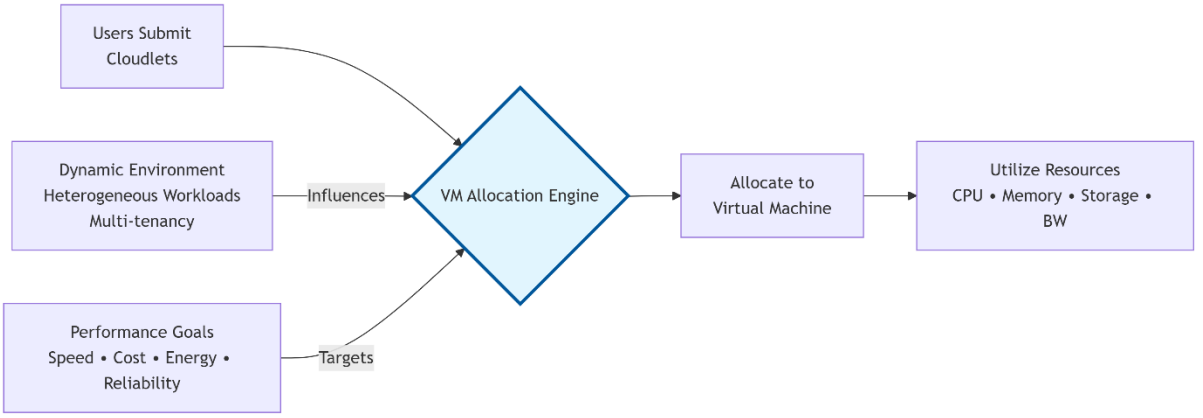

Fig. 1. Virtual Machine Allocation in Cloud Computing Architecture

VM allocation significantly influences the major performance metrics, which include execution time, throughput, energy consumption, operational cost, and reliability. Approaches to allocation may be **static**, based on predetermined workload characteristics, or dynamic, adjusting in real time, response to fluctuate system states [9]. Dynamic approaches increasingly employ predictive analytics, intelligent optimization, and learning-based algorithms to enhance decision-making accuracy and system responsiveness.

2.2. Fairness in VM Allocation

Fairness is a critical principle in multi-tenant cloud environments, where diverse users and tasks compete for shared computational resources. Fairness-oriented allocation mechanisms aim to prevent resource monopolization, reduce task starvation, enhance predictability, and improve user satisfaction. However, fairness introduces a design challenge: balancing equity with performance-oriented objectives such as throughput, cost efficiency, and energy savings [10-11]. Several fairness paradigms have been employed in VM allocation, these include proportional, max-min, and equity.

2.2.1. Proportional Fairness

Allocates resources in proportion to predefined weights such as priority level, subscription tier, or task importance as shown in Figure 2. This approach ensures that users receive resources relative to their contribution or expected service level [12].

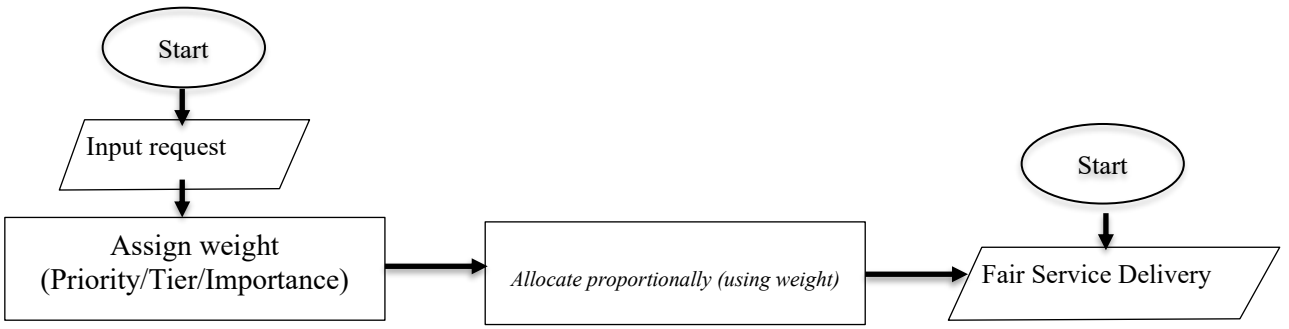

Fig. 2. Proportion Fairness allocation scheme

2.3. Taxonomy of Fairness Paradigms in VM Allocation

2.3.1. Proportional Fairness Foundations and Operational Adaptations

Early research in the 2020 epoch concentrated primarily on establishing robust economic and auction-based theoretical foundations for proportional fairness. Li et al. (2021) pioneered an innovative cloud auction framework where proportional equity was mathematically formulated as a constrained utility maximization problem under strict tenant budget boundaries [13]. Their methodology introduced a novel "virtual credit" allocation policy that scales proportionally with user-submitted bids, thereby enforcing truthful bidding strategies (incentive compatibility) while simultaneously preserving global system equity. This economic lens provided a market-driven justification for proportional resource distribution, proving that cloud providers could seamlessly implement multi-tiered service levels while maximizing global revenue streams.

Shifting the paradigm from purely economic abstractions toward operational and green-computing efficiency, Salami et al. (2020) integrated proportional fairness constraints into energy-aware scheduling frameworks within heterogeneous data centers [14]. Their architectural innovation lied in weighting resource allocation factors by combining tenant priority indices with server-specific thermal and energy-efficiency coefficients. Under this schema, high-priority, computationally intensive tasks were preferentially routed to highly efficient hardware nodes. This dual-criterion optimization framework represented a significant departure from traditional single-dimensional fairness models, successfully injecting environmental sustainability considerations into what had conventionally been treated as an isolated Quality of Service (QoS) problem.

Concurrently, the rapid proliferation of edge and fog computing infrastructures prompted Kumar et al. (2022) to adapt proportional fairness models for latency-sensitive applications operating within localized IoT environments [11]. Recognizing that rigid, static proportionality could inadvertently degrade strict deadline compliance, they engineered a temporal, deadline-aware proportional fairness algorithm. Within this architecture, allocation weights are dynamically re-calibrated in real time based on task urgency and residual laxity. This introduction of a temporal dimension marked a critical evolutionary phase in cloud scheduling, shifting fairness from a static resource constraint to an agile, state-responsive mechanism capable of balancing long-term inter-user equity with short-term, deterministic latency boundaries for Internet of Things (IoT) applications.

2.3.2. Max–Min Fairness and Starvation Mitigation Protocols

The core philosophy of Max-Min fairness revolves around maximizing the minimum resource share allocated across competing tasks, prioritizing the most underserved or resource-deprived entities to systematically eliminate extreme starvation. Although highly equitable, a recognized drawback of this paradigm is its tendency to diminish aggregate data center resource utilization. To mitigate this performance trade-off, several researchers have introduced hybrid structural enhancements to the classical Max-Min logic.

Raeisi-Varzaneh et al. (2021), for instance, proposed an Advanced Max-Min task scheduling framework tailored for multi-tenant cloud ecosystems, optimizing both economic cost-effectiveness and algorithmic throughput [15]. Their primary contribution was a dual-objective optimization architecture that directly embedded a cost-aware heuristic into the foundational Max-Min allocation loop. By dynamically adjusting task-to-VM mapping matrices, the algorithm jointly minimizes maximum makespan and financial execution costs. This methodology preserves the starvation-prevention guarantees inherent in Max-Min scheduling while noticeably elevating physical hardware utilization via real-time budget feedback loops. Their empirical benchmarks demonstrated that this cost-integrated variant significantly outperformed traditional baseline models, including standard Min-Min, classic Max-Min, and Shortest Job First (SJF) configurations, across metrics of waiting time and throughput.

Addressing the complexities of multi-resource environments, Zhang et al. (2022) targeted the pervasive challenge of resource fragmentation, a phenomenon where naive Max-Min execution severely misallocates secondary assets [16]. They formulated the "Dominant Resource with Max-Min Fairness" (DRF-MM) model, which extends classical Dominant Resource Fairness by uniformly applying Max-Min equilibrium vectors across both dominant and non-dominant hardware components. This formulation guarantees that users with highly divergent resource consumption

profiles (e.g., CPU-bound vs. memory-bound workloads) receive structurally equitable treatment based on their specific bottleneck constraints. The DRF-MM framework serves as a vital mathematical bridge between proportional utility models and strict Max-Min principles in highly heterogeneous cloud topologies. As distributed cloud systems continue to drift toward specialized accelerator-driven hardware, Max-Min fairness remains a critical algorithmic pillar for guaranteeing basic resource access virtualization.

2.3.3. Equity Fairness and Adaptive Predictive Modeling

Distinct from purely mathematical or economic distribution rules, Equity Fairness is deeply rooted in psychological, behavioral, and organizational socio-economic frameworks, such as Adams' Equity Theory. This paradigm prioritizes a user's perceived balance of the allocation process, incorporating historical infrastructure usage, computational effort, and past structural under-provisioning into the current scheduling decision matrix [17]. Resource orchestration under this umbrella underscores the necessity of designing adaptive, market-oriented strategies that resolve highly diverse service demands without eroding global QoS bounds or provider cost-efficiency.

Providing a foundational baseline for this utility-driven cloud paradigm, Buyya et al. (2008) pioneered the vision of cloud infrastructures operating similarly to traditional public utilities. Their seminal work advocated for market-driven resource management frameworks capable of balancing user-centric SLA expectations against provider-side financial and computational risks. By championing early virtualization technologies and exploring inter-cloud federation paradigms, their research highlighted the critical need for convergence among competing IT deployment models to achieve macro-level equitable resource distribution [18].

To transition these conceptual equity theories into operational deployments, empirical and predictive models have been widely explored to guide proactive resource provisioning. Islam et al. (2012) developed advanced artificial neural networks (ANN) and linear regression formulations capable of executing adaptive resource management, validating that predictive analytics drastically improve scheduler responsiveness under highly fluctuating, non-stationary demand signals [19]. Similarly, Singh et al. (2025) engineered an adaptive resource allocation framework tailored specifically for mobile cloud computing (MCC) environments, aiming to maximize cumulative system rewards by optimizing resource provisioning under severe local hardware and network constraints [20].

Collectively, these historical efforts emphasize that integrating predictive analytics, evolutionary multi-objective optimization, and market-driven utility models is essential for constructing a fair, responsive, and robust resource allocation ecosystem capable of satisfying the multi-faceted demands of contemporary cloud environments.

2.4. Optimization Paradigms in VM Allocation

Mathematical optimization forms the cornerstone of effective and sustainable Virtual Machine (VM) allocation strategies. Because optimal resource mapping across multi-dimensional cloud infrastructures is inherently NP-hard, deterministic brute-force exploration is computationally prohibitive for real-time systems. To navigate this algorithmic complexity, researchers have engineered a diverse spectrum of optimization models. As illustrated in contemporary cloud literature, these methodologies can be broadly taxonomized into heuristic policies, metaheuristic frameworks, and exact mathematical programming approaches.

2.4.1. Heuristic and Rule-Based Strategies

Heuristic approaches utilize deterministic, predefined rules such as First-Fit (FF), Best-Fit (BF), and Round-Robin (RR) to deliver fast, low-overhead resource allocation decisions. While computationally efficient and easily implementable within production hypervisors, these static policies often yield highly suboptimal solutions when deployed in large-scale, heterogeneous multi-tenant environments due to their lack of adaptive state-responsiveness [21].

Nevertheless, a substantial body of literature emphasizes the strategic refinement of heuristic frameworks to optimize VM placement, with a primary focus on thermal management and physical resource utilization. For instance, Feng et al. (2021) introduced an innovative heat-recirculation-aware VM placement heuristic that mitigates data center cooling energy consumption by explicitly modeling internal localized thermal dynamics [22]. Similarly, Hormozi et al. (2022) focused on server consolidation heuristics that leverage residual capacity defragmentation techniques to maximize energy efficiency while sustaining steady Quality of Service (QoS) metrics [23]. Furthermore, the foundational First-Fit (FF) heuristic remains a subject of active research; Mukhopadhyay & Tewari (2024) demonstrated that when properly integrated into cloud bin-packing frameworks, the FF policy provides a reliable baseline for macro-level energy conservation and active host management [24].

Recent technological advancements have begun integrating predictive machine learning models directly into classical heuristic engines to overcome their static limitations. A prominent example is the VTGAN architecture developed by Maiyza (2025), which couples generative prediction networks with the Power Best Fit Decreasing (PBFD) heuristic [25]. This framework proactively identifies and excludes physical hosts projected to undergo impending overload states, successfully optimizing energy usage without inflating SLA violation rates. Parallel to this, the predictive look-ahead strategies analyzed by Çağlar & Altılar (2022) utilize dynamic look-ahead heuristics to allocate virtualized infrastructure. By anticipating future capacity demand vectors over a sliding temporal horizon, these hybrid models drastically minimize unnecessary VM migrations and improve aggregate data center energy profiles [26].

2.4.2. Metaheuristic Frameworks and Global Search Mechanisms

Metaheuristic algorithms provide robust, stochastic global search capabilities that are exceptionally well-suited for resolving multi-objective, non-linear optimization challenges within dynamic cloud environments. The most widely adopted paradigms include:

Genetic Algorithms (GA): Explicitly model biological evolution to traverse complex, discontinuous search surfaces utilizing iterative selection, crossover, and mutation operations.

Particle Swarm Optimization (PSO): Leverages swarm intelligence principles to iteratively refine VM scheduling configurations by balancing localized individual velocity vectors with collective global coordinates.

Ant Colony Optimization (ACO): Simulates path-finding behavior based on artificial pheromone trails to determine optimal topological routing and task-to-host allocation trajectories.

While metaheuristics regularly achieve near-optimal solutions on complex surfaces, they require rigorous hyperparameter tuning and often introduce non-trivial computational latency, which can hinder their applicability in strict, real-time cloud operations [27].

To address these performance trade-offs, several recent investigations have targeted the coupled nature of VM management. Alsadie (2021) developed the Multi-Objective Dynamic VM Allocation (MDVMA) metaheuristic framework, engineered specifically to manage the joint complexity of concurrent task scheduling and VM placement when scaling to massive, scale-free cloud task matrices [28]. Similarly, Alboaneen et al. (2021) proposed a hybridized metaheuristic methodology that unifies scheduling pipelines with underlying VM assignment models, highlighting that integrated structural optimization significantly improves cluster-wide throughput compared to isolated scheduling loops [29].

From a comparative perspective, the comprehensive review by Singh et al. (2021) provided an analytical benchmarking of diverse metaheuristic algorithms deployed for heterogeneous task scheduling [30]. Their taxonomy evaluated algorithmic efficiency under multi-criteria constraints, illustrating the continuous evolution of swarm and evolutionary models. Additionally, the systematic review by Panwar et al. (2022) highlighted the efficiency of micro-genetic algorithms (\mu\text{-GA}) and specialized bio-inspired metaheuristics, validating their role as highly effective, low-overhead mechanisms for energy-aware VM consolidation and carbon footprint reduction in hyper-scale cloud data centers [31].

2.4.3. Deterministic Mathematical Optimization Formulations

In contrast to stochastic or rule-based methods, exact mathematical programming and proportional allocation models provide formalized, deterministic closed-form expressions for cloud resource distribution. Proportional resource allocation techniques compute absolute CPU, memory, and bandwidth shares utilizing mathematically defined weighted priority ratios, enabling highly transparent, reproducible, and verifiable scheduling decisions.

To formalize this mechanism under standard multi-tenant operational constraints, the baseline proportional allocation share *S_i* for a specific virtual machine instance *i* competing within a shared physical host capacity can be mathematically formulated as follows:

$$CPU_i = \frac{w_i}{\sum_{j=1}^{n} w_j} * C_{total} \quad (1)$$

Where:

= CPU share for task i CPU_i
= weight or contribution factor by task iw_i
= total CPU capacityC_{total}

The execution time can be represented as :

$$T_i = \frac{L_i}{CPU_i} \quad (2)$$

Where:

= execution time for task $iCPU_i$
= workload length (in MI)L_i

These models form the mathematical foundation for fairness-aware proportional allocation frameworks such as OPEM.

3. RESEARCH METHODOLOGY AND SYSTEMATIC REVIEW PROTOCOL

This systematic review implements a rigorous, evidence-based methodology designed in strict accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. By systematically identifying, evaluating, and synthesizing fairness-aware, optimization-driven, and hybrid Virtual Machine (VM) allocation models published between January 2020 and March 2025, this protocol establishes a reproducible and transparent foundation. This structured approach serves to map the contemporary cloud scheduling landscape and contextualize the positioning of the Optimized Proportional Equity Model (OPEM) within current architectural trends.

3.1. Research Design

The adoption of a Systematic Literature Review (SLR) paradigm is explicitly motivated by the exponential growth of cloud federation literature and the historically fragmented, secondary treatment of fairness within conventional VM orchestration research. To guarantee methodological rigor, transparency, and reproducibility which are highly scrutinized in computer science and software engineering surveys the review sequence carefully traces a formalized pipeline encompassing research question formulation, multi-database query execution, dual-stage eligibility screening, and qualitative data extraction.

3.2. Research Questions (RQs)

To establish a definitive analytical scope and guide the data extraction phase, this systematic review was governed by four primary research questions:

RQ1 (Taxonomy): What distinct Virtual Machine (VM) allocation architectures and models have been proposed within Infrastructure-as-a-Service (IaaS) cloud environments between 2020 and 2025?

RQ2 (Operational Fairness): How do existing multi-tenant cloud allocation frameworks mathematically operationalize, balance, and enforce fairness among competing users?

RQ3 (Optimization Techniques & Metrics): Which optimization paradigms (heuristic, metaheuristic, or exact mathematical methods) are most prevalently deployed, and what specific performance metrics (e.g., energy, makespan, cost) do they target?

RQ4 (Limitations & Gaps): What are the primary structural limitations, trade-offs, and open challenges identified in contemporary fairness-aware and optimization-driven cloud scheduling frameworks?

3.3. Data Sources and Search Strategy

A comprehensive, multi-disciplinary electronic search was executed across leading Scopus-indexed digital libraries to ensure the retrieval of high-impact, peer-reviewed literature. The targeted search repositories included IEEE Xplore, ScienceDirect (Elsevier), and the broader Scopus indexing database.

The search strings were constructed using advanced Boolean logic (AND/OR operators) to pair primary keyword clusters representing the domain, mechanism, and objectives:

\text{Query} = (\text{"cloud computing"} \lor \text{"IaaS"} \lor \text{"data center"}) \land (\text{"VM allocation"} \lor \text{"virtual machine placement"} \lor \text{"task scheduling"}) \land (\text{"fairness"} \lor \text{"proportional equity"} \lor \text{"optimization"})

To isolate the most modern technological paradigms and non-stationary cloud scheduler designs, the temporal boundaries of the query window were strictly confined to studies published between January 2020 and March 2025.

3.4. Eligibility Criteria (Inclusion and Exclusion)

To filter the initial search results and preserve a high-quality, methodologically sound corpus, a rigid set of eligibility criteria was established. Table 1 defines the explicit boundaries applied during the selection phase.

Table 1. Methodological Inclusion and Exclusion Criteria

Inclusion Criteria (IC)	Exclusion Criteria (EC)
• Peer-reviewed journal articles, conference proceedings, or book chapters published between Jan 2020 and Mar 2025.	• Non-peer-reviewed white papers, preprints (e.g., arXiv), extended abstracts, or editorial reviews.
• Studies focusing directly on VM allocation, virtualized resource distribution, or macrotask scheduling within IaaS platforms.	• Studies focusing exclusively on low-level hardware design, physical networking protocols, or cloud security without scheduling links.
• Frameworks that explicitly model or evaluate algorithmic fairness, proportional equity, or optimization-driven scheduling.	• Publications falling outside the designated 2020–2025 temporal horizon.
• Research presenting formal mathematical models, algorithmic pseudocode, or quantitative empirical performance evaluations.	• Studies lacking sufficient methodological depth, clear mathematical formulation, or validated experimental results.

3.4. Screening and Selection Pipeline

The initial pool of raw literature retrieved from the digital repositories underwent a systematic, dual-stage screening workflow to minimize selection bias.

Title and Abstract Screening: In the first phase, titles and abstracts were independently audited to exclude explicitly irrelevant studies, duplicate entries, and out-of-scope applications.

Full-Text Eligibility Assessment: In the second phase, the remaining candidate manuscripts were subjected to an exhaustive full-text evaluation. This step verified that each paper provided formal algorithmic configurations, robust evaluation metrics, or empirical validations regarding multi-tenant equity.

Through this rigorous distillation sequence, a final core corpus of 15 seminal papers was isolated for comprehensive qualitative and quantitative synthesis. This curated final selection captures the essential methodological divisions of modern cloud scheduling, encompassing rule-based heuristics, stochastic metaheuristics, deterministic proportional mathematical equations, and multi-objective hybrid frameworks. These 15 baseline papers form the foundational empirical evidence compiled in this review, enabling a structured benchmarking of state-of-the-art architectures. The complete end-to-end literature search, screening tiers, and final extraction counts are visually codified in the standard PRISMA flow diagram illustrated in Figure 3.

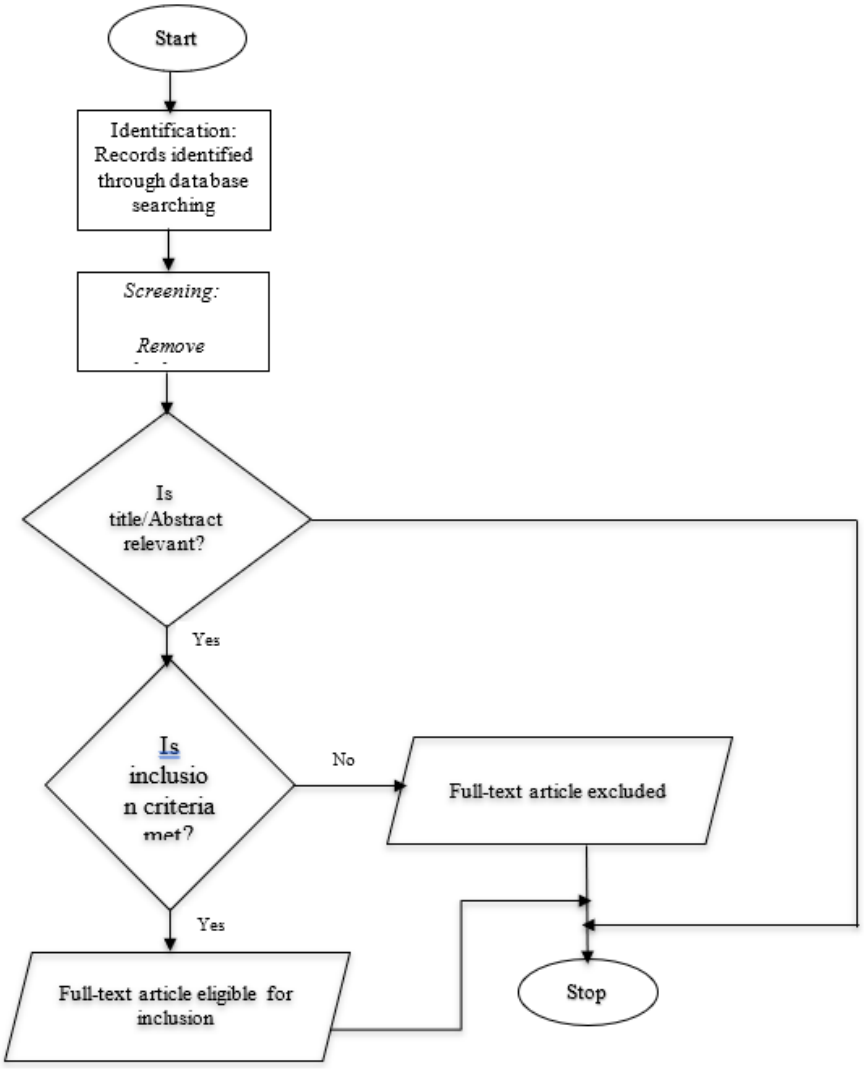

Fig. 3. Graph of the literature search and selection process

After the selection of the core papers, a systematic data extraction process was undertaken to enable a consistent comparative analysis. Some key information was employed from each study, including the author(s) and publication year, the primary allocation methodology (e.g., heuristic, meta-heuristic, mathematical, hybrid), the specific fairness considerations and metrics employed, and the performance metrics used for evaluation (e.g., execution time, cost, throughput, makespan), along with key findings and limitations. A tabular synthesis was employed to systematically compare these methods, highlight trends, and identify research gaps. This structured approach ensured the subsequent analysis could effectively examine how fairness and optimization objectives are integrated, the quantitative evaluation of resource allocation, and the practical implications for cloud environments.

3.5. Comparative Analysis of VM Allocation Models

The comparative analysis focuses on three dominant model categories: fairness-driven, optimization-driven, and hybrid fairness–optimization approaches. Evaluation is conducted across multiple dimensions, including performance efficiency, fairness assurance, computational complexity, scalability, and applicability to real-world cloud environments.

Such a multi-dimensional comparison is essential because VM allocation decisions often involve trade-offs between system-wide efficiency and equitable resource distribution in multi-tenant infrastructures.

3.5.1. Fairness-Driven Models: Ensuring Equity at a Cost

Fairness-driven models prioritize equitable resource distribution, often drawing on principles from economics and game theory. The primary strength of this paradigm is its explicit focus on preventing resource starvation and ensuring predictable performance for all users, particularly in multi-tenant environments.

Methodologies and Contributions: A prominent technique is the proportional share fairness model, as seen in the work of Singh (2023), where VMs are allocated based on predefined priority weights. This approach ensures that a high-priority task consistently receives a larger share of resources relative to its demand [9]. Other studies, such as Ishola et al. (2022), extended this concept by incorporating historical usage, creating an "equity-based" scheduler that rewards past resource contribution, thereby fostering a fairer long-term ecosystem [32].

3.5.2. Optimization-Driven Models: Pursuing Peak Efficiency

In stark contrast, optimization-driven models focus on maximizing system-centric metrics such as throughput, energy efficiency, and cost reduction, often treating user-centric fairness as a secondary concern.

Methodologies and Contributions: This category is dominated by meta-heuristic algorithms. Studies by Abdel-Basset et al. (2021) utilized hybrid whale optimization and grey wolf optimizer, respectively, to navigate the complex search space of VM-to-host mapping, achieving significant improvements in energy savings and load balancing compared to traditional first-fit or round-robin algorithms [3]. Similarly, Li et al. (2023) applied deep reinforcement learning (DRL) to learn dynamic provisioning policies that maximize provider profit while adhering to service level agreements [13].

Performance and Limitations: The strength of these models is their unparalleled performance in raw efficiency metrics. They consistently report superior results in reducing energy consumption by up to 20% and improving throughput by 15-30% over baseline methods. The critical weakness, however, is their potential to create performance polarization.

As Cohen (2025) observed, without explicit fairness constraints, these models can allocate resources in a way that benefits a majority of tasks at the severe detriment of a minority, violating principles of equitable quality of service (QoS) [33].

3.5.3. Hybrid Models: The Emerging Consensus

Hybrid model came up as a result of the limitations of siloed approaches, the most impactful recent research falls into the hybrid category. These models explicitly formulate VM allocation as a multi-objective problem, seeking a pareto-optimal balance between fairness and efficiency.

Methodologies and Contributions: The primary technique involves multi-objective optimization (MOO). Wang et al. (2024) developed a fairness-aware multi-objective particle swarm optimization (FA-MOPSO) that integrated a proportional fairness index directly into the particle's fitness function alongside makespan and cost [34].

Performance and Limitations: Hybrid models represent the current state-of-the-art, demonstrating that the fairness-efficiency trade-off is manageable. They typically achieve 90-95% of the efficiency of pure optimizers while maintaining high fairness scores. The main challenge is their increased complexity; designing and tuning a multi-objective function or a lexicographic hierarchy is non-trivial and can lead to longer convergence times for meta-heuristics.

Recent literature reflects a growing emphasis on fairness-oriented algorithms in VM allocation, particularly within multi-tenant cloud environments where diverse workloads compete for limited resources. Fairness considerations, once treated as secondary to throughput and efficiency, are now recognized as essential optimization objectives in modern VM allocation research because it prevents starvation of low-priority tasks, ensure predictable performance across user groups, and improve equity during periods of resource contention. Contemporary studies now adopt proportional, priority-aware, and equity-based mechanisms that create more transparent and balanced allocation outcomes.

However, despite this growing interest, several research gaps persist. A notable limitation is the absence of standardized fairness metrics. Studies rely on diverse indicators such as Jain's Fairness Index, Max–Min fairness, proportional deviation scores, utility-based metrics, and custom equity formulations. This inconsistency makes cross-study comparison subjective and hinders the establishment of a unified benchmark for evaluating fairness-oriented models. Additionally, many existing works continue to rely heavily on simulated environments like CloudSim, using small or synthetic workloads that fail to capture the complexity of real-world cloud infrastructures. Without evaluations based on large-scale, trace-driven, or production workloads, the real-world applicability and robustness of proposed algorithms remain uncertain.

A further concern is the limited provision of theoretical guarantees and scalability analyses. Only a few studies offer formal proofs, worst-case bounds, or convergence analyses, especially among metaheuristic and fairness-driven approaches. As a result, properties like stability under dynamic loads, predictability of allocation outcomes, and scalability to thousands of tasks and VMs remain insufficiently explored.

In a nutshell, the literature reveals a significant opportunity for the development of hybrid fairness–optimization frameworks. Existing fairness-centric models often improve equity at the expense of efficiency, while optimization-driven models maximize performance but overlook disparities among users or tasks. Integrating these approaches into unified multi-objective frameworks would enable systems to embed fairness constraints directly into optimization formulations, adapt trade-offs dynamically based on workload conditions, and maintain stable, predictable performance under fluctuating demand. Advancing such hybrid strategies remains a promising direction for future research, with strong potential for real-world adoption in commercial cloud environments.

4. PROPOSED OPEM AND ITS POSITION IN FAIR VM ALLOCATION

The Optimized Proportional Equity Model (OPEM) is proposed as a unified, fairness-aware VM allocation framework that advances the current landscape of resource management in multi-tenant cloud environments. While it builds on established proportional allocation principles, OPEM contributes novel conceptual value through its integration of fairness, performance, and cost into a single deterministic decision architecture. Unlike heuristic or meta-heuristic approaches that rely on approximation, stochastic exploration, or heuristic tuning, OPEM provides

mathematically transparent, reproducible allocation outcomes, offering a structured pathway for equitable and performance-conscious resource distribution.

At the core of OPEM is the principle that fairness should not be treated as an afterthought but embedded directly into the allocation mechanism. The model operationalizes fairness through contribution weights that go beyond traditional static priorities. These weights may encode multidimensional characteristics such as SLA class, user subscription tier, historical under-provisioning, latency sensitivity, or fairness compensation factors. This design elevates proportional allocation from a simple ratio-based rule to a richer fairness representation capable of capturing real-world heterogeneity in user demands and service obligations. Through this formulation, OPEM directly addresses one of the major research gaps: the absence of a standardized and interpretable fairness mechanism that can be applied consistently across studies.

A defining strength of OPEM is its deterministic structure. Meta-heuristic schedulers such as GA, PSO, ACO, and hybrid AI optimizers often produce near-optimal but non-reproducible outcomes because their search processes rely on randomness, convergence tuning, or domain-specific heuristics. These methods also introduce substantial computational overhead and require careful parameter calibration. In contrast, OPEM offers a transparent mapping from contribution weights to allocated MIPS, execution time, and processing cost, providing clear traceability for performance evaluation and system auditing. This determinism makes OPEM suitable as a baseline fairness model for benchmarking, a role that existing literature has struggled to fill due to inconsistent fairness metrics and algorithmic opacity.

OPEM is further positioned as a response to the growing demand for multi-objective reasoning in resource allocation. By explicitly linking fairness (proportional equity), performance (execution time), and cost (processing charges), the model establishes a fairness–performance–cost pipeline that enables systematic trade-off analysis. This integrated perspective bridges the persistent divide between fairness-driven research often criticized for reducing system throughput and optimization-centric models, which frequently overlook equity among competing tasks. In this way, OPEM advances the field by demonstrating how fairness can be mathematically encoded without abandoning efficiency or cost-awareness.

The research significance of OPEM also lies in its extensibility. The model provides a modular structure that can be incorporated into hybrid fairness–optimization frameworks, applied alongside AI-based schedulers, or expanded to include energy-awareness, deadline constraints, or dynamic weight adjustment under changing workloads. Its adaptability makes it relevant for real-world deployment, where cloud providers must justify allocation policies, reduce SLA violations, and maintain predictable performance amidst diverse and unpredictable workloads.

Overall, OPEM is positioned as a conceptually rigorous and practically deployable fairness framework that addresses key research gaps lack of standardized fairness metrics, limited interpretability of existing schedulers, and insufficient integration of equity with performance and cost. By offering a deterministic, mathematically grounded, and fairness-centered approach, OPEM advances the discourse on cloud resource allocation and provides a new foundation upon which future fairness-aware and multi-objective scheduling models may be developed.

The flowchart in Figure 4 illustrates the Optimized Proportional Equity Model (OPEM) for VM allocation follows a structured, deterministic workflow designed to embed fairness directly into scheduling decisions. The process begins when multi-tenant users submit VM requests, from which OPEM extracts multi-dimensional contribution weights, such as SLA class, subscription tier, historical under-provisioning, latency sensitivity, and fairness compensation factors. These weights are normalized, aggregated, and fed into a proportional allocation formula that maps each user's weight share to a concrete resource assignment (MIPS, execution time, and cost). A fairness-threshold check then validates the allocation against predefined equity metrics; if the threshold is not met, the model adjusts the weights or re-balances the queue in a feedback loop until fairness is satisfied. Once approved, the schedule is executed, and every allocation outcome is transparently logged providing a reproducible record for auditing and future fairness adjustments. This closed-loop, rule-based architecture ensures that OPEM delivers consistent, interpretable, and fair resource distribution without relying on stochastic or heuristic optimization.

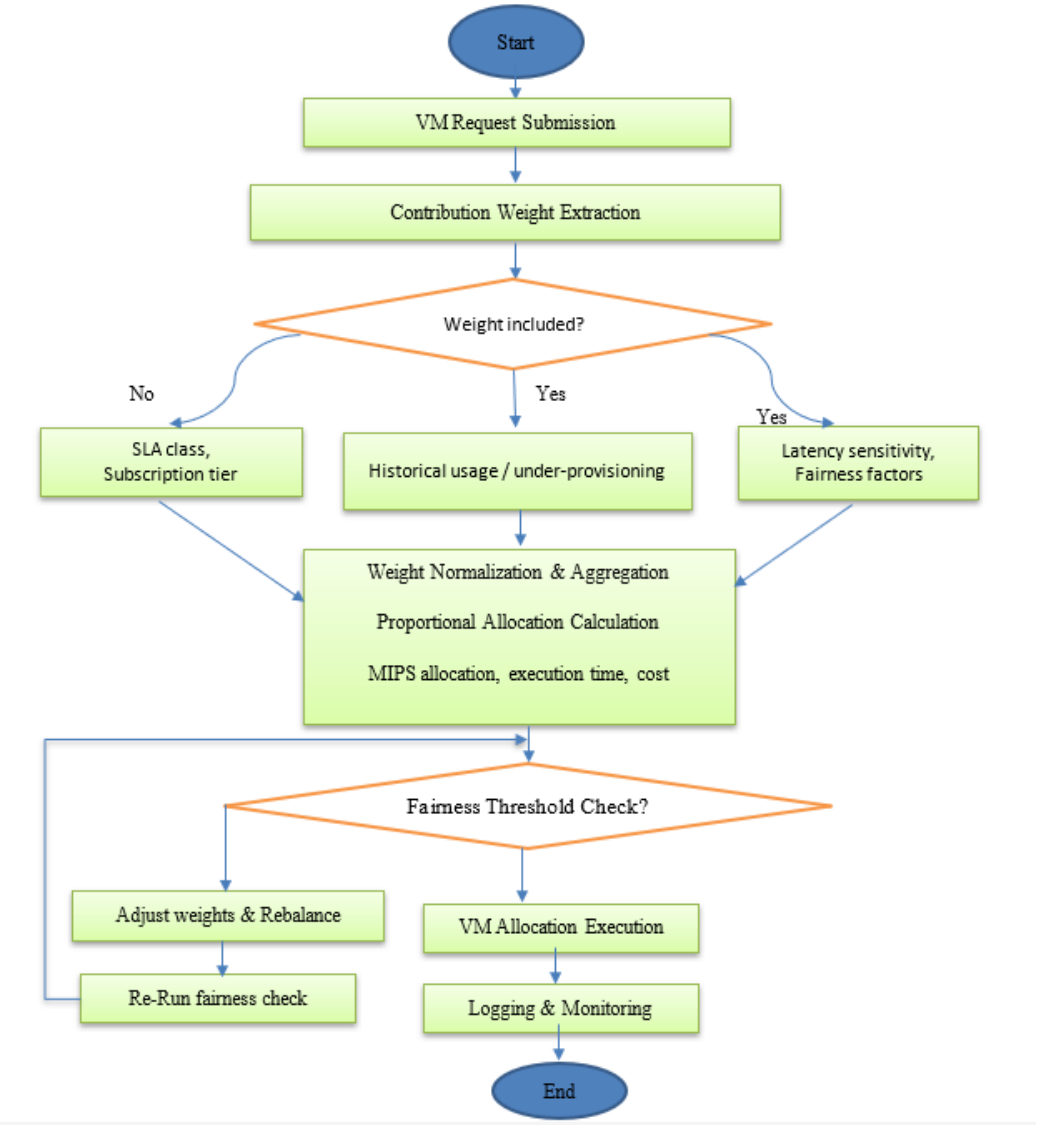

Fig. 4. OPEM Workflow

5. CONCLUSION

This review examined fifteen key studies published between 2020 and 2025 on virtual machine (VM) allocation, with a focus on fairness-oriented and optimization-driven models. The analysis reveals a clear trend in the research community toward multi-objective and fairness-aware resource allocation approaches. Whereas earlier works primarily targeted throughput, energy efficiency, or makespan, recent studies increasingly recognize that equity is essential for achieving stable and efficient cloud performance.

The reviewed literature highlights several important observations. First, fairness and optimization should be viewed as complementary objectives rather than mutually exclusive, particularly in modern heterogeneous cloud environments. Second, there remains no unified metric for fairness, which complicates the benchmarking and comparison of allocation algorithms. Third, while optimization techniques, including heuristics, meta-heuristics, and mathematical programming continue to evolve, many approaches lack scalability validation using realistic

workloads. Finally, there is a growing need for models that balance fairness, efficiency, and system dynamics to address the complexities of multi-tenant cloud infrastructures.

Overall, the body of research demonstrates substantial progress in both fairness-aware and optimization-based VM allocation strategies. At the same time, persistent methodological gaps remain, particularly regarding standardization of metrics, workload realism, and theoretical guarantees, indicating clear directions for future investigation.

Declaration

We acknowledge that we used ChatGPT to enhance the academic writing of our manuscript while ensuring the originality and integrity of our work.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Rana, N., Jeribi, F., Khan, Z., Alrawagfeh, W., Ben Dhaou, I., Haseebuddin, M., & Uddin, M. (2024). A systematic literature review on contemporary and future trends in virtual machine scheduling techniques in cloud and multi-access computing. Frontiers in Computer Science, 6, 1288552. https://doi.org/10.3389/fcomp.2024.1288552

[2] Sun, W. (2020). An optimal resource allocation scheme for virtual machine placement. AIMS Mathematics, 5(3), 2290-2306. https://doi.org/10.3934/math.2020256

[3] Abdel-Basset, M., Mohamed, R., Abouhawwash, M., Chakrabortty, R. K., & Ryan, M. J. (2021). EA-MSCA: An effective energy-aware multi-objective modified sine-cosine algorithm for real-time task scheduling in multiprocessor systems: Methods and analysis. Expert systems with applications, 173, 114699. https://doi.org/10.1016/j.eswa.2021.114699

[4] Barbalho, H., Kovaleski, P., Li, B., Marshall, L., Molinaro, M., Pan, A. & Menache, I. (2023). Virtual machine allocation with lifetime predictions. Proceedings of Machine Learning and Systems, 5, 232-253.

[5] Gu, H., Wang, J., Yu, J., Wang, D., Li, B., He, X., & Yin, X. (2023). Towards virtual machine scheduling research based on multi-decision AHP method in the cloud computing platform. PeerJ Computer Science, 9, e1675. https://doi.org/10.7717/peerj-cs.1675

[6] Li, J., & Yang, C. (2023). Hybrid machine learning and optimization approach for resource allocation in edge computing. Journal of Cloud Computing: Theory and Applications, 12(1), 28-45.

[7] Kumar, H., Garg, A., Gupta, A., & Agarwal, Y. (2025). A Hybrid Proactive And Predictive Framework For Edge Cloud Resource Management. arXiv preprint arXiv:2511.16075.

[8] Oyekanmi, E. O., Oladoja, I. P., & Omotehinwa, T. O. (2021). Improved RASA for load balancing in cloud computing environments. Asian Journal of Research in Computer Science, 12(4), 52-66. https://doi.org/10.9734/ajrcos/2021/v12i430294

[9] Singh, J., & Walia, N. K. (2023). A comprehensive review of cloud computing virtual machine consolidation. IEEE Access, 11, 106190-106209. https://doi.org/10.1109/ACCESS.2023.3314613

[10] Hosseini, H., & Schierreich, Š. (2025). The Algorithmic Landscape of Fair and Efficient Distribution of Delivery Orders in the Gig Economy. arXiv preprint arXiv:2503.16002.

[11] Kumar, S., Goswami, A., Gupta, R., Singh, S. P., & Lay-Ekuakille, A. (2022). A cost-effective and QoS-aware user allocation approach for edge computing enabled IoT. IEEE Internet of Things Journal, 10(2), 1696-1710. https://doi.org/10.1109/JIOT.2022.3210835

[12] Hamzeh, H. (2021). Fairness for resource allocation in cloud computing (Doctoral dissertation, Bournemouth University).

[13] Li, S., Huang, J., & Cheng, B. (2021). Resource pricing and demand allocation for revenue maximization in IaaS clouds: A market-oriented approach. IEEE Transactions on Network and Service Management, 18(3), 3460-3475. https://doi.org/10.1109/TNSM.2021.3085519

[14] Salami, B., Noori, H., & Naghibzadeh, M. (2020). Fairness-aware energy efficient scheduling on heterogeneous multi-core processors. IEEE Transactions on Computers, 70(1), 72-82. https://doi.org/10.1109/TC.2020.2984607

[15] Raeisi-Varzaneh, M., Dakkak, O., Fazea, Y., & Kaosar, M. G. (2024). Advanced cost-aware Max-Min workflow tasks allocation and scheduling in cloud computing systems. Cluster computing, 27(9), 13407-13419. https://doi.org/10.1007/s10586-024-04594-1

[16] Zhang, X., Li, J., Li, G., & Li, W. (2022). Generalized asset fairness mechanism for multi-resource fair allocation mechanism with two different types of resources. Cluster Computing, 25(5), 3389-3403. https://doi.org/10.1007/s10586-022-03548-9

[17] Tao, M., Hao, F., Wei, L., Zhi, H., Kuznetsov, S. O., & Min, G. (2024). Fairness-aware maximal cliques identification in attributed social networks with concept-cognitive learning. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2024.3445721

[18] Buyya, R., Yeo, C. S., & Venugopal, S. (2008). Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In 2008 10th IEEE international conference on high performance computing and communications (pp. 5-13). Ieee. https://doi.org/10.1109/HPCC.2008.172

[19] Islam, S., Keung, J., Lee, K., & Liu, A. (2012). Empirical prediction models for adaptive resource provisioning in the cloud. Future Generation Computer Systems, 28(1), 155-162. https://doi.org/10.1016/j.future.2011.05.027

[20] Singh, A. R., Sujatha, M. S., Kadu, A. D., Bajaj, M., Addis, H., & Sarada, K. (2025). A deep learning and IoT-driven framework for real-time adaptive resource allocation and grid optimization in smart energy systems. Scientific Reports, 15. https://doi.org/10.1038/s41598-025-02649-w

[21] Singhal, S., Gupta, N., Berwal, P., Naveed, Q. N., Lasisi, A., & Wodajo, A. W. (2023). Energy efficient resource allocation in cloud environment using metaheuristic algorithm. Ieee Access, 11, 126135-126146.

https://doi.org/10.1109/ACCESS.2023.3330434

[22] Feng, H., Deng, Y., Zhou, Y., & Min, G. (2021). Towards heat-recirculation-aware virtual machine placement in data centers. IEEE Transactions on Network and Service Management, 19(1), 256-270. https://doi.org/10.1109/TNSM.2021.3120295

[23] Hormozi, E., Hu, S., Ding, Z., Tian, Y. C., Wang, Y. G., Yu, Z. G., & Zhang, W. (2022). Energy-efficient virtual machine placement in data centres via an accelerated Genetic Algorithm with improved fitness computation. Energy, 252, 123884. https://doi.org/10.1016/j.energy.2022.123884

[24] Mukhopadhyay, N., & Tewari, B. P. (2024). Performance Analysis of a Virtualized Cloud Data Center through Efficient VM Migrations. In Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing (pp. 357-365). https://doi.org/10.1145/3675888.3676071

[25] Maiyza, A. I., Hassan, H. A., Sheta, W. M., Banawan, K., & Korany, N. O. (2025). VTGAN based proactive VM consolidation in cloud data centers using value and trend approaches. Scientific Reports, 15(1), 20133. https://doi.org/10.1038/s41598-025-04757-z

[26] Çağlar, İ., & Altılar, D. T. (2022). Look-ahead energy efficient VM allocation approach for data centers. Journal of Cloud Computing, 11(1), 11. https://doi.org/10.1186/s13677-022-00281-x

[27] 27--Ramamurthy, A., Pantula, P. D., Gharote, M. S., Mitra, K., & Lodha, S. (2021). Multi-objective Optimization for Virtual Machine Allocation in Computational Scientific Workflow under Uncertainty. In CLOSER (pp. 240-247). https://doi.org/10.5220/0010453302400247

[28] Alsadie, D. (2021). A metaheuristic framework for dynamic virtual machine allocation with optimized task scheduling in cloud data centers. IEEE Access, 9, 74218-74233. https://doi.org/10.1109/ACCESS.2021.3077901

[29] Alboaneen, D., Tianfield, H., Zhang, Y., & Pranggono, B. (2021). A metaheuristic method for joint task scheduling and virtual machine placement in cloud data centers. Future Generation Computer Systems, 115, 201-212. https://doi.org/10.1016/j.future.2020.08.036

[30] Singh, H., Tyagi, S., Kumar, P., Gill, S. S., & Buyya, R. (2021). Metaheuristics for scheduling of heterogeneous tasks in cloud computing environments: Analysis, performance evaluation, and future directions. Simulation Modelling Practice and Theory, 111, 102353. https://doi.org/10.1016/j.simpat.2021.102353

[31] Panwar, S. S., Rauthan, M. M. S., & Barthwal, V. (2022). A systematic review on effective energy utilization management strategies in cloud data centers. Journal of Cloud Computing, 11(1), 95. https://doi.org/10.1186/s13677-022-00368-5

[32] Ishola, O. M., & Oyedele, O. O. (2022). Fairness-aware resource allocation in cloud computing. Journal of Computer Science and Its Applications, 9(2), 1-8.

[33] Cohen, M. C., Miao, S., & Wang, Y. (2025). Dynamic Pricing with Fairness Constraints. Operational Research, 73, 3027-3043. https://doi.org/10.1287/opre.2023.0123

[34] Wang, W., Li, B., & Liang, B. (2014). Dominant resource fairness in cloud computing systems with heterogeneous servers. In IEEE INFOCOM 2014-IEEE Conference on Computer Communications (pp. 583-591). IEEE. https://doi.org/10.1109/INFOCOM.2014.6847983