Journal of Information Systems and Telecommunication
Vol.13, No.3, July-September 2025, 210-231
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Towards Energy-efficient Cloud Computing: A Review of Network-Aware VM Placement Approaches

Ali M. Baydoun[1*], Ahmed S. Zekri [2]

[1].Department of Mathematics & Computer Science, Beirut Arab University, Lebanon
[2] Department of Mathematics & Computer Science, Alexandria University, Egypt

Received: 03 Jan 2025/ Revised: 04 Sep 2025/ Accepted: 20 Aug 2025

Abstract
Cloud data centers (CDCs) have witnessed significant growth to meet the increasing demands of modern applications. However, this expansion has raised concerns regarding the environmental impact, energy requirements, and electricity costs associated with data centers. The network infrastructure, serving as the communication backbone of these data centers, plays a crucial role in their scalability, performance, cost, and, most importantly, energy consumption. This review provides meaningful perspectives and valuable insights into the state-of-the-art research regarding the problem of virtual machine placement (VMP), focusing on the network-aware energy efficiency aspects of data centers. It provides an overview of VM placement and presents a comprehensive survey of prominent VM placement algorithms from the existing literature. Additionally, a thematic taxonomy of network-aware algorithms is introduced, highlighting the key energy consumption metrics and presenting a new classification of VMP algorithms that considers datacenter network (DCN) topology, traffic patterns, communication patterns, and energy reduction strategies. Besides addressing pertinent research questions in this domain, this review summarizes the findings and suggests potential avenues for future research, guiding researchers in designing and implementing more effective and efficient network-aware VM placement algorithms that optimize energy consumption, improve network performance, and minimize migration costs.

Keywords: Cloud computing; VM placement; network-aware; Energy-efficient; Network architecture.

1- Introduction

Cloud computing is an internet-based technology that provides services without the need for physical infrastructure ownership. The cloud computing model is responsible for managing tens of data centers that manage computing applications and data storage. Cloud providers offer three service models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), with deployment models including public, private, community, and hybrid [1]. Virtualization is the key factor in cloud computing. It improves resource efficiency and reduces costs. The high energy consumption in data centers is a significant issue, especially with cooling equipment that consumes 80% of available energy [2].
In the cloud environment, virtual machine (VM) traffic can account for 50%-80% of total data center network traffic [3], motivating network-aware placement to minimize cross-rack hops and reduce energy consumption. In this field, most research focuses on optimizing resource utilization and power consumption to address cost-related challenges. Proper planning of the network architecture is very important as the number of VMs continues to rise and data centers and communication networks continue to expand. As cloud applications handle more data, inter-VM network bandwidth increases due to the high demand for bandwidth that heavily depends on network resources. This presents a challenge for cloud environments to strike a balance between energy efficiency and performance. Conserving energy through reducing network equipment could lead to a violation of service level agreements (SLAs) and degrade performance [4].
Why Network-Aware VM Placement Matters:
Despite growing efforts to optimize server energy use, the network infrastructure —comprising switches, routers, and links— remains a major yet often under-optimized contributor to overall energy consumption. What makes network-aware VM placement particularly compelling is its

✉ **Ali M. Baydoun**
al.baydoun@bau.edu.lb

dual impact: it not only reduces energy usage by limiting inter-rack communication and enabling low-power network states but also improves performance by lowering latency and congestion. These benefits become increasingly relevant as VM-to-VM communication dominates traffic patterns in modern data centers. As such, placement strategies must now evolve to consider network topology and traffic locality as primary optimization dimensions, not secondary concerns.

This paper explores several research questions related to network-aware VM placement in cloud data centers (CDCs). It begins by analyzing the key factors previously examined in this domain, such as initial VM placement and potential migrations, and their impact on network performance. The study then identifies the most effective metrics for evaluating the success of energy-efficient, network-aware VM placement algorithms, considering both resource utilization and network performance. Additionally, it investigates how the network topology within a data center affects overall power consumption and whether enhancing network power efficiency can influence the costs associated with VM migration.

This paper makes the following contributions to the field of energy-efficient, network-aware VM placement in CDCs:

- Taxonomy of Methodologies

We propose a novel taxonomy that systematically classifies existing network-aware VM placement approaches, highlighting each approach's underlying energy-efficiency mechanisms.

- Categorization of Existing Work

We analyze and categorize state-of-the-art algorithms based on key metrics —such as topology awareness, traffic patterns, and consolidation techniques— and evaluate their impact on overall energy consumption.

- Identification of Challenges

We pinpoint critical gaps in current research, most notably the lack of integration between VM placement strategies and dynamic network energy-saving techniques .

- Proposed Solutions

We suggest actionable solutions to address these challenges, including cross-layer optimization frameworks and topology-aware VM consolidation heuristics that co-locate high-traffic VMs to minimize network usage.

- Future Research Directions

We outline open problems and emerging trends; such as AI-driven placement and edge-cloud coordination; to guide future work in this area.

- Practical Resource for Researchers

We provide a structured reference for practitioners, showing how to balance network performance and power savings when designing new VM placement algorithms.

The remainder of this paper is organized as follows. Section 2 reviews existing surveys on network-aware VM placement. Section 3 presents an analysis of VM placement (VMP) algorithms. Section 4 introduces our taxonomy of network-aware, energy-efficient approaches. Section 5 discusses the limitations of today's research. Finally, Section 6 concludes with key takeaways and outlines precise future research directions aimed at helping both researchers and practitioners design VM placement strategies that minimize power usage without compromising network performance.

2- Landscape of Existing VMP Surveys

2-1- Overview of Prior Surveys Focus Areas

Several survey articles have previously explored VMP in cloud computing, addressing critical challenges in areas such as minimizing energy consumption, optimizing traffic routing, and ensuring resource allocation efficiency. These efforts span a wide range of algorithmic strategies, including heuristic algorithms, meta-heuristic optimization, dynamic workload balancing, and energy-aware scheduling. While individually rich in contributions, many of these surveys tend to focus on isolated dimensions of the VMP problem, often treating energy-efficiency and network-awareness as distinct objectives rather interdependent system constraints.

Although prior surveys cover individual hardware mechanisms—Dynamic Voltage and Frequency Scaling (DVFS) and Adaptive Link Rate (ALR) —or network-aware placement separately, no integrative framework treats these energy-saving techniques and network-sensitive parameters (traffic patterns, communication behavior, Datacenter Network (DCN) topology) as co-dependent.

- DVFS dynamically lowers a processor's supply voltage and clock frequency during light workloads to reduce power consumption.

- ALR reduces the data-link speed (or puts links into low-power idle modes) on underutilized network ports, saving significant switch and NIC energy but introducing variable latency when ramping back to full rate.

This deficiency limits the applicability of existing classifications in real-world CDCs where network usage and energy dynamics are deeply intertwined. Therefore, this review aims to bridge that gap by delivering a unified analytical lens that evaluates VMP strategies at the intersection of network topology, traffic behavior, and energy optimization—providing researchers and practitioners with a holistic foundation for future algorithmic developments.

2-2- Features and Gaps

Table 1 presents a multi-dimensional mapping of prior VMP surveys across several core features, highlighting areas of emphasis and omission in relation to network-awareness, energy-efficiency, and VM placement logic.

Table 1. Comparison of Existing Surveys on Network-Aware VM Placement Across Key Dimensions

Ref	Year	Placement & Migration	Traffic-Eng.	DCN Topology	Inter-VM/ VM→Storage	Comm. Pattern	Energy-Saving	Hardware-Based	Traffic-Based	Thermal Mgmt.	Perf. Impact	App Focus
[5]	2013	✗	✓	✗	✗	✗	✓	✓	✗	✓	✗	✗
[6]	2014	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗
[7]	2015	✓	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗
[8]	2014	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗
[9]	2014	✗	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗
[10]	2015	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✓
[11]	2015	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
[12]	2016	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
[13]	2020	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
[14]	2020	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗
[15]	2021	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✓
[16]	2023	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
[17]	2024	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓
[18]	2024	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
Our Work	2025	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

To further contextualize the strengths and omissions across surveys, Table 2 summarizes the primary focus of each reference and the most prominent gaps with respect to network-awareness and energy optimization.

Table 2. Most Prominent Gaps Across Reviewed Surveys.

Ref	Year & Venue	Primary Focus	Most Prominent Gaps (in Network-Aware Context)
[5]	2013, Cluster Computing	ALR and link-layer energy techniques	No VM placement or topology-aware placement; lacks traffic pattern integration
[6]	2014, ACM Computing Surveys	High-level energy-efficiency (DVFS, link sleep)	Algorithmic VM placement details missing; no explicit DCN topology analysis
[7]	2015, FGCS	Network-aware VM placement & migration	No link-layer ALR/DVFS inclusion; limited thermal considerations
[8]	2014, Computer Communications	DCN architectures & energy-aware routing	No VM consolidation or ALR integration; lacks detailed performance vs. energy metrics
[9]	2014, FGCS	Green DCN architectures taxonomy	Hardware-level focus; lacks VM-level dynamics or traffic/thermal overlays
[10]	2015, JNCA	Live VM migration & server consolidation frameworks	Limited network awareness (focuses on migration traffic); does not tie placement to topology or ALR
[11]	2015, IEEE CCGrid	General VM placement taxonomy	Does not explicitly cover network-energy techniques (ALR) or topology variations
[12]	2016, JNCA	Algorithm catalog (ILP, heuristics, metaheuristics)	Lacks network-energy integration; does not address dynamic traffic patterns
[13]	2020, JSC	Multi-objective VM placement	Does not integrate ALR or DCN topology; limited discussion of per-flow traffic metrics
[14]	2020, Kybernetes	Classification of VMP mechanisms in cloud	No explicit focus on link-layer energy or inter-VM traffic topology
[15]	2021, Computer Science Review	Multi-level consolidation (VM, container, etc.)	No focus on ALR or DCN topology; limited to consolidation trends
[16]	2023, The Journal of Computational Science and Engineering	Review of 7 energy-efficient VM placement strategies	General efficiency metrics; lacks deep integration of DCN traffic patterns or communication metrics
[17]	2024, Frontiers in Computer Science	ML-based VM scheduling techniques	Does not classify topologies or link-level policies; lacks VM clustering detail
[18]	2024, Telecommunication Systems	Phased VMC lifecycle review (PM→VM selection→placement)	Does not integrate link-layer energy or topology; focuses on VM phases without network-energy objectives
—	2025, TBD (Our Work)	Unified network-aware VMP taxonomy	Fills all gaps by integrating ALR, topology, traffic patterns, and energy/thermal considerations

While Table 1 and Table 2 provide a comparative overview of survey scopes, a deeper analysis of each work reveals further insights into thematic priorities and overlooked dimensions. As summarized in Table 2, the majority of prior surveys fail to integrate link-layer energy mechanisms, DCN topology constraints, and traffic-aware placement into a unified classification framework. This motivates the need for a closer, qualitative critique of each referenced study—highlighting what each survey addresses and, more importantly, how our work advances beyond them with a network-aware energy-efficient focus.

2-3- Critical Analysis

This subsection presents an evaluation of each major survey study on VMP published from 2013 through 2024, with a focus on their contributions to energy-efficient and network-aware strategies. For each referenced work ([5]-[18]), we describe the main idea of the survey, identify its strengths, and highlight gaps related to the intersection of communication patterns, topology constraints, and power efficiency. Such analysis has two goals: first, to document the advancement of the domain in the past ten years, and second, to show how most of these surveys fail to integrate all these aspects into a single framework. This subsection also serves to demonstrate how our proposed taxonomy explicitly addresses these multi-layered challenges by integrating network topology, traffic-awareness, and energy-aware mechanisms under a unified VM placement perspective. These observations establish the rationale for our integrated taxonomy, as elaborated in the following sections.

The survey [5] offer one of the foundational treatments of green networking by categorizing ALR techniques - dividing link-sleep policies (immediate vs. delayed wake) and link-rate scaling schemes- and by evaluating the IEEE 802.3az standard's potential to save nearly 0.9 TWh annually in large US data centers. Their strength lies in rigorously detailing how ALR can dynamically reduce link-layer power, from NICs up to aggregation switches. However, because their focus remains at the hardware and firmware level, they do not address how VM placement or migration strategies might leverage fluctuating link speeds or ALR states to optimize overall data center energy. Our survey fills this gap by explicitly integrating ALR considerations into the network-aware VM placement taxonomy, demonstrating how VM co-location based on communication affinity can complement hardware-level ALR to maximize energy savings.

The authors of [6] present a broad, multi-layer survey of energy-efficiency techniques in large-scale distributed systems, covering hardware-level approaches (DVFS, power modeling), server-level optimizations (VM consolidation, dynamic provisioning), and network-layer tactics (ALR, link-sleep, topology reconfiguration). Their work's strength is in demonstrating that up to 30−40% of a data center's energy can be consumed by its networking infrastructure, thus motivating holistic solutions, but lacks a taxonomy specific to VM placement. Our work fills this void by extending network-layer concerns into VM placement contexts, thereby illustrating how topology- and traffic-aware placement strategies interact with server and link energy dynamics.

The authors of [7] present a specialized taxonomy of network-aware VM placement and migration algorithms, classifying approaches based on problem formulation (ILP vs. heuristics), traffic awareness (static vs. dynamic), and objectives (minimizing inter-VM traffic, avoiding congestion, balancing network load) . They survey methods that co-locate high-traffic VM pairs -reducing inter-rack hop counts by roughly 30%. Although they excel in highlighting how inter-VM communication patterns drive placement, they do not incorporate link-layer ALR or DVFS as explicit dimensions in their classification, nor do they quantify the impact of particular DCN topologies on overall energy consumption. Our survey extends their work by embedding these network-aware placement algorithms within a broader framework, explicitly incorporating DCN structure, traffic distribution patterns, and link utilization characteristics into placement decision-making.

Authors in [8] provides a focused survey on architectures and energy efficiency in data center networks. It covers DCN topologies (FatTree, VL2) and green techniques like link adaptation and component shutdown. However, it lacks granularity in VM-level policies. Our review complements this by showing how such architectural designs can be better utilized when paired with VM placement that respects traffic distribution and energy states, offering specific placement criteria that leverage topology-induced communication cost differences.

The authors in [9] conducted a comprehensive survey on Green Data Center Networks (DCNs), focusing on energy-efficient architectures (electrical, optical, hybrid), traffic management, and performance monitoring. While their work extensively covers network-level energy optimization techniques like ALR and topology-aware resource consolidation, it does not systematically integrate VM placement strategies with network energy efficiency. This separation weakens the applicability of their insights for practical scheduling decisions. This work integrates their hardware-level insights into VM placement taxonomy, connecting traffic profiles and server locality to DCN energy states.

The authors of [10] deliver a deep examination of live VM migration and server consolidation frameworks, categorizing bandwidth-optimization techniques (block-level and file-level deduplication, delta compression, dynamic rate limiting), storage-checkpoint approaches, and consolidation triggers (CPU/memory thresholds vs. predictive models). Their strength is in quantifying

✉ Ali M. Baydoun
al.baydoun@bau.edu.lb

migration downtime, total transfer time, and migration energy overhead across dozens of tools (e.g., Xen pre-copy, KVM post-copy, RDMA-accelerated). They also survey DVFS-enabled consolidation policies that reduce CPU power during migration windows. However, they do not incorporate network-awareness beyond minimizing migration traffic; specifically, they do not explore how VM selection and placement decisions could optimize for inter-VM communication patterns. In contrast, our survey extends their consolidation framework by explicitly modeling migration and placement objectives that minimize both compute and network power.

The work in [11] propose a five-axis taxonomy for VM placement —spanning optimization objectives (power, performance, network, reliability), workload models (batch, enterprise, web, HPC), constraints (QoS, SLA, affinity), problem formulations (ILP, CP, heuristics, metaheuristics), and placement modes (static vs. dynamic). They provided researchers with an early, systematic way to navigate the VM placement literature. Nonetheless, their taxonomy does not explicitly integrate network-layer energy techniques such as ALR or discuss how specific DCN topologies shape algorithmic design. Our work builds on their multi-dimensional approach by DCN topology —thus mapping each placement algorithm onto a richer, network-aware energy context, and explicitly correlating traffic patterns with link-power-saving opportunities.

Survey [12] compile an extensive algorithm-centric overview of VM placement techniques, grouping them into exact ILP/MIP formulations, multi-objective nonlinear programming, bin-packing heuristics (e.g., First-Fit Decreasing, Best-Fit Decreasing), coalition- and graph-theory methods (e.g., Hungarian algorithm), and evolutionary metaheuristics (GA, PSO, ACO, SA, BBO) . They evaluate each category in terms of scalability, solution quality, and runtime, concluding that metaheuristics predominate for large data centers. However, their survey omits any discussion of network-aware energy techniques or DCN topology. In our work, we situate each algorithm class within a unified, network-aware framework that specifies how each network metric studied influence performance and energy outcomes, thereby providing practical guidance on selecting placement strategies based on the communication structure of the workload.

In their study [13], the authors deliver a comprehensive multi-objective taxonomy for IaaS VM placement, distinguishing between single-objective (power only) and multi-objective (power and network, power and QoS) methods, and between operation modes (offline vs. online), while also noting emerging challenges such as AI/ML-based placement and edge-cloud integration. However, they do not unify ALR or DCN topology into their taxonomy. Our survey builds upon their multi-objective perspective by adding a network-energy dimension, including communication-aware cost functions and DCN-aware co-location policies.

The survey [14] provides a comprehensive overview of VMP mechanisms in cloud environments by systematically categorizing approaches into static and schemes. Their strength lies in rigorously detailing the mapping algorithms, selection criteria, and resource-utilization impacts across 40 carefully filtered studies. However, because their focus remains at the process level (static vs. dynamic) and general algorithmic families, they do not analyze how network-aware strategies, thermal considerations, or renewable-energy profiles influence VMP decisions. Our survey fills this gap by explicitly integrating these concerns, by enabling sustainability-oriented VM allocation guided by real-world infrastructure constraints.

The work described in [15] resent a comprehensive survey of data center consolidation in cloud computing systems, with a significant portion dedicated to VM-level consolidation techniques —examining threshold-based host selection, VM selection heuristics, and consolidation-driven energy models for CPU and memory utilization. Their strength lies in synthesizing a wide range of VM consolidation algorithms−ranging from simple first-fit and best-fit heuristics to more advanced ILP and metaheuristic formulations−and in highlighting how VM consolidation can reduce the number of active hosts and, consequently, overall energy consumption. However, although they touch on VM migration overhead, they do not incorporate network energy considerations nor analyze how specific data center topologies influence consolidation decisions. Our survey extends their VM-level focus by embedding each consolidation algorithm within a network-aware framework, explicitly showing how inter-VM traffic patterns interact with placement heuristics to maximize combined compute and network energy savings, resulting in more holistic and topology-sensitive consolidation strategies.

The authors of [16] present a concise survey of seven energy-efficient VM-placement algorithms in cloud data centers, covering load-balancing heuristics, metaheuristic methods, queuing-based models, simulation-driven approaches, static placement schemes, hybrid strategies, and predictive control techniques. Their work's strength lies in clearly summarizing each algorithm's core mechanism and practical applicability, but it lacks a systematic taxonomy and quantitative comparison—particularly omitting network-layer energy management. Our survey fills this void by introducing a comprehensive, multi-dimensional taxonomy and detailed comparison tables that explicitly integrate network- and thermal-aware dimensions into VM placement strategies, bridging infrastructure constraints with algorithm design.

The authors of [17] conduct a systematic literature review (SLR) of VM-scheduling studies, categorizing them into three principal methodologies —traditional, heuristic, and

meta-heuristic— and rigorously charting their problem formulations, performance metrics, and simulation environments. Their strength lies in applying a clear SLR protocol to distill trends and challenges across a broad corpus. However, because their taxonomy is organized solely around algorithmic families and general scheduling parameters, it omits network-aware energy management considerations. Our survey fills this void by introducing dedicated network- and thermal-awareness in the VM-placement classification, highlighting the impact of link-power state models and topology-aware routing in placement evaluation.

Authors of [18] offer a systematic overview of VM Consolidation (VMC) by describing the three fundamental phases -(1) Physical Machine (PM) detection, (2) VM selection, and (3) VM placement- and classifying works according to their problem formulation (ILP, heuristic, metaheuristic), constraint sets (SLA, affinity, resource capacities), and objective functions (power minimization, network traffic reduction, cost, SLA violation) . Their major contribution is the clear, phase-by-phase breakdown of VMC, which helps researchers identify algorithmic gaps in each subproblem. Still, although they recognize "minimizing network traffic" as one possible objective, they do not assess the role of DCN topology. In contrast, our survey embeds topology-aware metrics directly into the VMP decision model—linking traffic routing patterns, bandwidth bottlenecks, and link power profiles with placement granularity.

2-4- Motivation Toward a Network-Energy-Aware VMP Taxonomy

Building on the limitations identified, we now motivate the need for a more unified taxonomy that explicitly links energy and network metrics in VM placement.

This paper addresses these gaps by:

- Providing an integrated taxonomy covering both network and energy optimization.
- Categorizing and analyzing methods across heuristic, meta-heuristic, ML, and hybrid strategies.
- Highlighting topological and communication-aware metrics used in real deployments.
- Incorporating recent advancements (2022–2025) including RL-based, and graph-theory-informed VMP strategies.

In summary, the existing body of survey work demonstrates valuable insights into VM placement challenges, yet lacks a unified treatment that integrates network topology, communication behavior, and energy efficiency within a cohesive evaluation framework. These gaps underscore the importance of establishing a systematic classification of VMP strategies, not only to contextualize existing methods but also to lay the groundwork for deeper, network-aware taxonomic analysis.

In the following section, we present a general classification of VM placement approaches, categorizing them by strategic objectives, optimization techniques, infrastructure considerations, and workload profiles — all of which form the foundation for the specialized taxonomy introduced in Section 4.

Early research prioritized server-side optimization because DCNs were heavily overprovisioned and per-flow traffic metrics were not readily exposed to hypervisors. Moreover, combining server and network objectives created complex multi-objective problems, and only with the advent of SDN-based telemetry [7] did network-aware placement become both feasible and attractive.

2-5- Bibliometric Overview

To assess the scholarly rigor of our survey corpus, we first defined precise selection criteria—keywords related to virtual machine placement, inclusion of peer-reviewed articles from reputable publishers, and exclusion of non-technical reports or non-English sources. We then executed systematic searches across Scopus and Web of Science using Boolean combinations of "virtual machine placement," "cloud data center," and "energy efficiency," restricting results to publications between 2009 and 2025. After de-duplication and application of our inclusion/exclusion rules, 80 references remained for analysis. Table 3 summarizes the distribution of these works by their SCImago Journal Rank quartile and lists the corresponding reference numbers. Table 4 shows the temporal breakdown of the references into 2009–2018, 2019–2021, and ≥ 2022 periods. Together, these tables provide a clear picture of both the scholarly rigor and the evolution of the field over time.

Table 3. Distribution of survey references by SCImago journal rank quartile.

Quartile	Count	References
Q1	21	[6], [8], [9], [10], [12], [22], [26], [33], [37], [38], [44], [49], [52], [54], [60], [62], [69], [72], [73], [78], [85]
Q2	17	[5], [13], [15], [17], [21], [24], [30], [31], [39], [40], [45], [50], [63], [66], [76], [79], [83]
Q3	8	[14], [18], [35], [36], [47], [57], [70], [74]
Q4	5	[2], [23], [34], [53], [80]
N/A	34	[1], [3], [4], [7], [11], [16], [19], [20], [25], [27], [28], [29], [32], [41], [42], [43], [46], [48], [51], [55], [56], [58], [59], [61], [64], [65], [67], [68], [71], [75], [77], [81], [82], [84]

All Quartiles are taken from the latest SCImago data (2024).

Conference proceedings, book chapters, standards, preprints, and other non-journal venues are marked N/A.

Table 4. Distribution of survey references by publication period (2009–2018, 2019–2021, ≥ 2022).

Date Range	Count	Reference Numbers
2009–2018	38	[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [25], [27], [32], [42], [53], [54], [55], [56], [58], [59], [60], [62], [64], [65], [66], [67], [68], [70], [71], [72], [73], [74], [75], [77], [78], [80]
2019–2021	21	[13], [14], [15], [26], [29], [31], [34], [36], [37], [40], [41], [43], [46], [47], [49], [50], [51], [57], [69], [76], [85]
2022 and after	26	[16], [17], [18], [19], [20], [21], [22], [23], [24], [28], [30], [33], [35], [38], [39], [44], [45], [48], [52], [61], [63], [79], [81], [82], [83], [84]

3- VM Placement Classification

This section reviews VM-level placement techniques in IaaS clouds. While container orchestration (e.g. Kubernetes, Docker Swarm) and serverless paradigms are reshaping resource management, they lie outside our VM-centric focus. For multi-level consolidation spanning VMs and containers, we refer readers to [15].

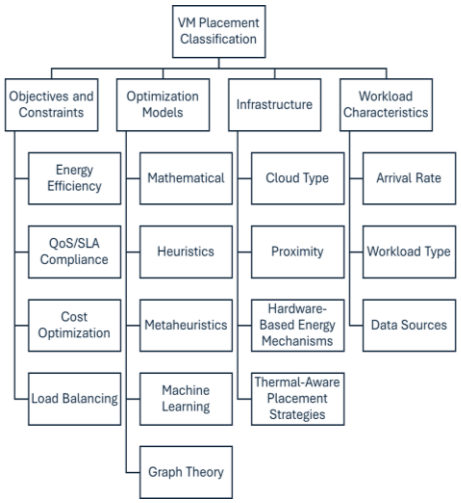

Fig. 1. VM Placement Classification

To establish a foundation for network-aware taxonomic refinement, we first present a generalized classification of VMP strategies. This section categorizes the existing approaches through four essential questions as shown in Fig.1—Why place?(Objectives), How to place?(Methods), Where to place?(Constraints), and What is being placed?(Workload)—each representing a pillar of modern VMP design. It is important to note that many studies do not fit in a single category. Instead, authors often formulate their placement strategies using a combination of objectives, methods, and constraints, leading to intentional overlap across these classification boundaries. This multidimensional design reflects the complex, real-world trade-offs that cloud service providers must manage.

3-1- Placement Objectives & Constraints (Why Place?)

A- Energy Efficiency

Energy efficiency is a foundational objective in VM placement, targeting both server-side and network-side power reductions. At the server level, strategies such as consolidation and intelligent VM distribution aim to reduce the number of active physical PMs. On the network side, minimizing inter-VM communication distance—by placing frequently interacting VMs closer within the topology—reduces switch and link utilization.

The Energy Efficient VM Placement (EE-VMP) model proposed in [19] demonstrated remarkable improvements, reducing power consumption by up to 56.89% and the number of active servers by 37%, while enhancing resource utilization by over 64%. These results underscore the potential of topology-aware consolidation combined with server optimization. However, the algorithm depends on accurate traffic matrices, which are rarely available in real time.

Similarly, an Active Energy-Efficient Placement method [20] achieved average energy reductions of 21.2% compared to the First Fit baseline. This highlights the efficacy of lightweight heuristic decision-making when real-time adaptability is needed, particularly in large-scale public clouds. However, its simplicity ignores inter-VM traffic patterns, potentially increasing cross-rack communication. Thus, Active Placement is attractive for compute-heavy, low-communication workloads but falls short when inter-VM latency and bandwidth must also be managed.

For dynamic workloads, the MOEA/D-based placement method proposed by [21] provides a more nuanced multi-objective balance. It simultaneously minimizes energy usage and overload risks, ensuring QoS compliance while maintaining performance efficiency under load. This approach is especially valuable in heterogeneous cloud environments with fluctuating demand, although it comes at the cost of higher computational complexity. That said, it adds significant computational cost. Choosing MOEA/D is advisable when offline tuning is acceptable and runtime overhead is secondary to multi-objective precision; otherwise, one should reject it in favor of faster approximation methods.

In [22], authors propose an algorithm designed to jointly minimize the energy consumption of both servers and network devices. The algorithm incorporates traffic awareness by co-locating highly interactive VMs and selecting physical paths with minimal energy costs. Their results demonstrated 11.4% reduction in total energy consumption, up to 22.3% reduction in network power

usage, and significant improvement in VM-to-VM communication efficiency. This method shows how intelligent mapping of traffic-heavy VMs to proximity-aware PMs can lower the utilization of aggregation and core switches, reducing link activation and routing overhead, yet the solution assumes that accurate traffic matrices are available prior to placement—a condition not always feasible in real-time cloud workloads.

B- QoS/SLA Compliance

Guaranteeing Quality of Service (QoS) and minimizing Service Level Agreement (SLA) violations are crucial objectives in VM placement. Overlooking these considerations can result in degraded user experience, financial penalties, and reduced provider reputation—especially in multi-tenant cloud infrastructures operating under tight availability thresholds.

The work in [23] introduced a utilization-aware VM placement policy that anticipates workload demands and avoids host overloading. By forecasting CPU trends and limiting consolidation aggressiveness, the method minimizes SLA violation time per active host while maintaining consolidation efficiency. However, reliance on CPU-only forecasting neglects network congestion effects during live migrations, potentially shifting bottlenecks to oversubscribed links. Moreover, the threshold-based decision logic may misfire under sudden workload spikes, degrading performance.

In [24], the authors proposed an Energy and QoS-aware VM placement algorithm (EQVMP) tailored for IaaS cloud environments. Their work integrates host energy modeling with service availability constraints, using a hybrid scheduling policy to minimize SLA violations. Experimental results show that EQVMP achieves lower energy consumption compared to baseline algorithms like RR and FF, while improving response time and reducing SLA violations, particularly under high-demand scenarios. Nevertheless, EQVMP's energy model abstracts away fine-grained network costs, and its rule-based availability checks introduce additional scheduling latency.

In a broader context, In [25], authors developed a multi-domain SLA management model incorporating a Generic SLA Manager (GSLAM) linked with OpenStack. Their approach models SLA violations and penalties across the IaaS, PaaS, and SaaS layers. The AV/AVL algorithms they introduce maintain availability above 99.99% and reduce penalty propagation across domains by controlling live migration overhead and optimizing host selection. While this multi-layer perspective improves service-level economics, the framework's orchestration complexity and cross-layer coordination overhead pose significant scalability challenges.

C- Cost Optimization

Cost-efficient VM placement remains a critical challenge in cloud infrastructures, especially in geographically distributed data centers where energy prices, carbon taxes, and renewable availability vary significantly. The work in [26] proposed a renewable- and carbon-aware VM allocation model that minimizes electricity costs and CO_2 emissions by dynamically placing VMs across data centers based on green energy availability, carbon intensity, and electricity prices. Their system integrates DVFS techniques and dynamic workload balancing, optimizing both cooling and server power usage. This work implicitly touches on network-related cost considerations by analyzing the carbon footprint and latency constraints tied to inter-data center VM placement and container communication, making it relevant to network-aware resource allocation. However, the method presumes reliable, low-latency energy pricing and renewable forecasts, which may not be universally available; it also overlooks performance impacts of inter-site VM migrations, risking degraded QoS for latency-sensitive workloads.

Similarly, in [27] authors designed a power and cost-aware placement strategy using a fuzzy decision model that simultaneously considers power consumption, electricity costs, and resource utilization. Their strategy yields measurable cost benefits under stable network conditions but omits dynamic bandwidth pricing and incurs significant overhead from fuzzy parameter tuning.

D- Load Balancing

Effective load balancing in virtual machine placement ensures even distribution of tasks across physical resources, which reduces processing delays, prevents host overloading, and maintains optimal system throughput. Load imbalance can lead to resource contention, degraded performance, or energy inefficiencies, particularly in high-density cloud environments.

In [28], a hybrid metaheuristic approach combining Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Artificial Bee Colony (ABC) is introduced to improve load distribution. This tri-hybrid method leverages the strengths of each algorithm: ACO's path-finding accuracy, PSO's global exploration, and ABC's exploitation of good solutions. The algorithm dynamically reallocates workloads among VMs based on current utilization, minimizing makespan and improving response time. Simulation using CloudAnalyst showed that the hybrid strategy significantly reduced average response time and execution time, outperforming classical load balancing algorithms like DLMA and IDLBA. Despite these gains, the combined algorithm entails high computational complexity, complex parameter calibration, and limited scalability under dynamic workloads.

Authors of [29] proposed the Min-Max Exclusive VM Placement (MMEVMP) strategy designed for scientific data environments, where workloads are data-intensive and disk I/O becomes a performance bottleneck. Unlike conventional CPU-centric methods, MMEVMP considers both disk bandwidth and CPU utilization to minimize SLA violations and reduce system operating costs. The algorithm dynamically avoids hosts likely to face disk saturation by analyzing historical usage patterns and applying adaptive time-based thresholds. Their experiments using a lightweight CloudSim version showed that MMEVMP achieved lower SLA violation rates while keeping energy consumption within acceptable bounds. However, the approach depends on accurate historical I/O profiling and neglects real-time network traffic patterns, potentially shifting bottlenecks to the network layer.

3-2- Optimization Models (How to place?)

Optimization approaches to VMP can be categorized into distinct yet overlapping models, each with advantages tied to performance, scalability, and adaptability to multi-objective goals. These include mathematical models, heuristic methods, metaheuristics, and learning-based approaches.

A- Mathematical Optimization

The work [30] presents a Multi-Objective Integer Linear Programming (MOILP) model for optimal VM placement, addressing resource utilization in CDCs. Although MOILP offers a rigorous mathematical framework for balancing conflicting objectives, its computational complexity grows exponentially with problem size. When applied to scenarios involving thousands of VMs and PMs, this leads to long solution times and excessive resource demands—rendering MOILP impractical for real-time or highly dynamic cloud environments. Even with enhancements like Tabu Search acceleration, solver runtimes extend beyond acceptable limits for dynamic cloud environments.

This paper [31] introduces mixed-integer programming (MIP) models for virtual machine placement that embed disk anti-colocation constraints—ensuring no physical disk hosts more than one virtual disk from the same VM—to optimize resource allocation in datacenters. MIP formulation may involve trillions of variables and/or constraints for large datacenter and therefore can't solve VMP optimally within acceptable time.

Optimization-based VM placement approaches offer mathematically rigorous formulations that guarantee optimality under well-defined constraints. These methods are especially suitable for precision-critical environments where deterministic outcomes are essential. Their ability to handle multiple objectives simultaneously (e.g., minimizing energy while balancing load and respecting hardware constraints) is a significant strength not easily replicated by heuristics or learning-based methods.

However, the computational cost of solving such models grows exponentially with problem size, making them impractical for large-scale cloud infrastructures [32]. Incorporating network-related constraints—such as inter-VM bandwidth demands, link capacities, or communication topologies—further increases the complexity. Even when advanced solvers or acceleration techniques are used, real-time placement decisions remain out of reach for anything beyond small- to medium-scale scenarios.

These approaches are also highly sensitive to changes in input parameters or constraints. A minor modification in workload demand or infrastructure policy may require full model regeneration and resolution, limiting their responsiveness to dynamic or elastic cloud environments. Furthermore, despite their theoretical strength in modeling energy consumption or network utilization, embedding such metrics into optimization formulations significantly delays solver convergence.

In terms of scalability, scenarios with fewer than 500 VMs are well-suited to these methods. On the other hand, large-scale, dynamic, or latency-sensitive platforms—such as public clouds or edge computing environments—are poorly matched due to the models' inability to respond within strict time constraints.

This type of optimization is best suited for offline placement in private clouds with stable demand, small-scale deployments where optimality justifies runtime, and regulated environments requiring strict constraint handling (e.g., security or compliance-based placement). But they perform worse with rapidly scaling public clouds, edge scenarios with latency bounds, and dynamic workloads requiring frequent re-optimization.

B- Heuristics

Heuristic methods are variants of bin-packing and greedy placement. They offer rapid, scalable approximations for the VM placement problem. Use simple, rule-based strategies (e.g. First-Fit, Best-Fit Decreasing [33])). These algorithms sort VMs by one or more dimensions (such as CPU demand or traffic volume) and assign each VM to the "best" host in linear or near-linear time.

GMPR [34] is a greedy placement algorithm that first ranks PMs by power efficiency to minimize the number of active hosts, then sequentially reduces resource imbalance and slack. In simulations on synthetic workloads and Amazon EC2 traces, GMPR achieves average savings of 1.91% in energy consumption and 16.18% in resource wastage versus state-of-the-art methods yet overlooks bandwidth costs.

Hybrid Best-Fit (HBF) [35] extends the classic Best-Fit heuristic by running three VM-ordering schemes (original, ascending size, descending size) and selecting the allocation with the lowest total energy. HBF consistently outperforms

both Best-Fit and Best-Fit Decreasing with minimal additional computation, but without addressing network proximity.

Heuristic-based VM placement approaches are widely used for their speed, simplicity, and scalability, making them particularly effective in large-scale datacenter environments where rapid decisions are essential. Techniques such as First-Fit and Best-Fit Decreasing achieve linear or near-linear time complexity (O(n log n)), enabling quick allocation of VMs with minimal computational overhead. Rule-based strategies, like sorting VMs based on CPU demand or traffic volume, are easy to implement and impose very little runtime cost. These methods are especially well-suited for static or predictable workloads.

However, the main limitation of heuristic approaches lies in their tendency to optimize single dimension while neglecting critical factors like network traffic. As a result, they often perform poorly in multi-objective optimization scenarios that require balancing energy consumption, latency, and SLA compliance. Their static nature also makes them not suitable for dynamic or unpredictable environments, where workload patterns change rapidly and real-time re-optimization is essential. While their computational efficiency remains a major strength, this speed frequently comes at the cost of placement accuracy compared to more adaptive metaheuristic or learning-based methods.

In terms of scalability, heuristics perform well, handling high volumes of VM requests. They are ideal for environments where quick and frequent placement decisions are needed without deep optimization logic. However, their suitability for energy- and network-aware placement remains limited. Although variants like HBF reduce host-level energy consumption, they do not model dynamic power states or account for network bandwidth costs, resulting in potentially inefficient traffic patterns. Overall, heuristics are best reserved for static or predictable workloads —such as batch processing— or for initial placement stages before applying more adaptive optimization techniques. They are less appropriate for network-intensive applications, dynamic edge environments, or scenarios demanding multi-objective trade-offs.

C- Metaheuristics

Metaheuristic approaches, such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Genetic Algorithms (GA), Grey Wolf Optimization (GWO), and their hybrids; tackle VM placement as a multi-objective optimization problem, balancing energy consumption, resource utilization, and SLA guarantees.

For example, [36] propose a hybrid ACO–GWO that weaves in traffic-awareness to co-locate high-communication VMs, yielding up to 19.41% power savings and 10.72% bandwidth-utilization improvements over baseline algorithms.

[37] classify and critique a broad spectrum of nature-inspired metaheuristics−SA, PSO, GA, ACO, BBO, and hybrids—highlighting their strengths in exploration/exploitation balance but noting their general omission of communication costs.

The work [38] presents a hybrid GA–best-fit scheme that minimizes active PMs and resource wastage, characterizing VMs by CPU, RAM, and bandwidth.

Recently, the work [39] proposed the NCRA-DP-ACO algorithm, a network-, cost-, and renewable-aware ACO framework for energy-efficient VM placement across geographically distributed datacenters. Unlike previous metaheuristic solutions, this work introduces a dynamic Power Usage Effectiveness (PUE) model, real-time solar energy profiling, and carbon-aware cost modeling. By integrating environmental and economic factors into the multi-objective placement strategy, the algorithm achieved up to 18% energy savings and a 48% reduction in live migrations compared to baseline heuristics and metaheuristics. This approach demonstrates that incorporating sustainability-aware factors can significantly enhance placement decisions in large-scale cloud environments, addressing a critical gap often neglected in earlier VM placement studies.

Metaheuristics offer excellent pathways to near-optimal placement of VMs in multi-objective environment. They are capable of compromising among energy efficiency, SLA, and resource consolidation while covering a large solution space.

However, their performance heavily depends on proper parameter tuning, and poor configurations lead to suboptimal convergence. Moreover, most metaheuristics neglect traffic patterns or topology, and therefore require additional improvements for traffic- and communication-aware optimizations. Enhanced variants can improve network efficiency but require additional computational overhead.

Since these algorithms are iterative and population-based searches over multiple generations (denoted as t), they exhibit higher O complexity —$O(n^2 \times t)$, where n is the problem size and t is the number of iterations. This reflects a quadratic growth in computational cost with problem size, meaning convergence time increases significantly as the number of VMs scales. Nevertheless, these approaches remain effective for medium to large problem sizes.

These approaches are best suited for offline or semi-dynamic VM placement scenarios where computation time is not a concern. They excel in multi-objective optimization —balancing energy efficiency, performance, and cost—and are effective in sustainable cloud environments that require periodic reallocation. However, they are less ideal for low-

latency edge computing due to slower convergence rates, and they tend to underperform in highly dynamic or unpredictable workloads where rapid re-optimization is essential. For small-scale deployments, simpler heuristic methods are often more practical.

D- Machine Learning

Emerging AI-driven VM placement frameworks leverage predictive and adaptive techniques to anticipate demand, group workloads, and continuously learn optimal allocations. Workload Forecasting Models employ learning-based algorithms to predict future load patterns and proactively select hosts that balance energy consumption and SLA adherence.

Classification & Clustering approaches identify high-traffic VM pairs or hosts at risk of overload and refine placement heuristics; Finally, Reinforcement Learning optimizes VM placement by learning from interactions with the environment (servers, network, and workloads).

Workload Forecasting Models: The work [40] introduces a dynamic, learning-based scheme that continuously predicts per-VM resource-usage thresholds to drive proactive allocation and live migration decisions. The approach adapts to fluctuating loads by generating runtime data and training a hybrid model (combining swarm-inspired search with an ML classifier), thus improving SLA compliance, reducing migrations, and cutting energy compared to standalone bio-inspired or ML methods.

Classification & Clustering: Random Forests or K-means identify which VM pairs generate the most traffic, or which hosts are likely to become overloaded, refining heuristic weightings. LECC [41] — a multi-objective VM (and data) placement framework for geo-distributed clouds that jointly minimizes carbon emission cost, energy consumption, and WAN communication cost— embeds an intelligent ML module that is trained on historical energy, latency, and carbon-cost data to dynamically adjust its multi-objective weightings (carbon emission, energy, WAN cost) at runtime. Extensive simulations on synthetic and real (PlanetLab and EC2) traces demonstrate LECC's ability to reduce server energy and cut response latency compared to baseline methods.

Reinforcement Learning (RL): The work [42] proposes a fuzzy-based State-Action-Reward-State-Action (SARSA) reinforcement learning algorithm for optimal VM placement in CDCs, effectively reallocating VMs to minimize energy consumption and resource wastage while ensuring compliance with SLA and QoS demands during fluctuating workloads.

ML-based VM placement algorithms adapt better than static heuristics under workload variation and fast-changing user demands.

Yet, there do exist serious disadvantages. These algorithms need huge amounts of training data of almost perfect quality, and their predictive power degrades if they are not promptly retrained or adapted. Many approaches in ML tend to disregard network traffic behavior or the underlying topology, limiting their applicability in optimizing network energy consumption or communication latency. These models add a further computational overhead and convergence delays: For instance, clustering methods scale at $O(n^3)$, while deep-learning techniques demand tremendous GPU/CPU resources [43].

Lastly, scalability becomes an issue: whereas the bigger data can continue to scale the ML model, on the other side, training and inference times increase with the size of the problem. Some solutions —distributed or federated learning— can help but introduce synchronization and convergence delays.

Network- and energy-aware suitability, and also optimization, are still primary concerns of most of these ML-based solutions. Advanced architectures like GNNs can integrate network topology into their learning workflow, but these models are computationally costly and thus seldom used. Without explicitly modeling bandwidth consumption or link-layer power states, ML-based placements may underperform when communication and geo-distribution dominate the environment [44].

ML-based VM placement algorithms are more suited to dynamic and large-scale cloud environments with regular patterns of workload and good availability of historical data [45]. However, their applicability is limited in real time or latency-sensitive deployments, where response has to be immediate. They also fail in environments where the workloads are unpredictable or rapidly changing.

E- Graph Approaches

Graph-theoretic VM placement models represent PMs/ VMs as graph nodes, with edges encoding constraints like inter-VM traffic or power costs. By applying community-detection or graph-partitioning algorithms, they co-locate highly communicative VMs −minimizing network hops and energy consumption.

The algorithm in [46] uses a graph-coloring algorithm that models VMs as graph vertices and inter-VM traffic volumes as weighted edges, then iteratively "colors" (assigns) and merges vertices to minimize both network overhead and server power use. Their method batches VM migrations to keep high-traffic groups co-located and decommission underutilized hosts. Extensive simulations across hierarchical datacenter topologies demonstrate that GCA halves link saturation and outperforms single-migration schemes by up to 65% in network-overhead reduction.

Authors in [47] propose a two-phase, graph-theoretic VM placement strategy tailored for data-intensive cloud applications. They first model the datacenter as a complete weighted graph —vertices are hosts, edges carry a networking-cost metric combining link saturation and hop count. In Phase 1, a fuzzy inference system ranks racks by

free resources and intra-rack traffic, and a linear program selects the smallest set of "close" racks with low uplink load. In Phase 2, the Traffic-Distance-Balanced (TDB) greedy algorithm uses the graph's weighted adjacency matrix to iteratively pick hosts minimizing total inter-host networking cost. This approach unifies capacity and communication in a single graph framework, ensuring high host utilization while keeping over 80% of traffic rack-local and halving link saturation compared to flat heuristics.

Despite clear advantages in topology-aware grouping, graph methods incur $O(n^3)$ complexity and often require full-network snapshots, impractical for frequent re-optimizations.

Despite their strength in encoding traffic and topology awareness, these methods come with high computational costs. Algorithms for community detection, graph partitioning, and coloring frequently exhibit $O(n^3)$ complexity, which becomes a bottleneck in large or fast-evolving systems [46].

Another limitation lies in their reliance on static or snapshot-based views of the network state. To remain effective, graph-based models require up-to-date global topology and traffic matrices —information that is difficult to capture or maintain in real time without imposing significant monitoring and re-computation overhead. Additionally, integrating these specialized algorithms into existing cloud controllers or schedulers remains a challenge due to their architectural differences.

From an energy and network efficiency perspective, graph-theoretic strategies outperform heuristic or ML-based approaches in minimizing communication overhead and active link utilization. However, this often comes at the expense of higher host-level energy consumption when traffic-based clustering leads to VM consolidation on less energy-efficient machines. While the network energy savings are clear, careful balance is required to avoid increasing overall compute energy due to suboptimal host selection. These algorithms are suitable for communication-intensive workloads with predictable traffic patterns (e.g., Hadoop), and hierarchical (or structured) data centers where intra-rack traffic locality is critical. However they perform poor with: real-time architectures with rapidly shifting traffic flows, edge and fog computing scenarios with strict latency constraints, and hyperscale public clouds (>10,000 VMs) where $O(n^3)$ complexity is unjustified [48].

Summary and Comparative Insights

While each VM placement strategy category—mathematical optimization, heuristics, metaheuristics, machine learning, and graph theory—has distinct merits, they also exhibit significant trade-offs in terms of computational complexity, scalability, and suitability for energy- and network-aware objectives. Mathematical optimization-based methods provide provable optimality for small-scale problems but are intractable for real-time or large deployments. Heuristic methods are fast and scalable but fail to consider complex objectives or traffic metrics. Metaheuristics deliver near-optimal results and support multi-objective optimization, yet often suffer from parameter sensitivity and long runtimes. ML approaches bring adaptability and prediction to dynamic environments but are data-hungry and rarely embed network topology or energy metrics explicitly. Graph-theoretic models excel at topology-aware co-location but incur high computational costs and require complete snapshot data. As summarized in Table 5, selecting an appropriate placement strategy requires balancing complexity, performance goals, and environmental context, especially when aiming to reduce both host and network energy consumption.

and uniform PUE values, supporting static or light dynamic heuristics [49]. However, placement strategies risk creating network congestion at the rack level if VM affinities are misestimated and lack resilience against localized failures or flash crowd events. Centralized placements suit applications with consistent workload distributions but should be augmented with fault-tolerance and burst-handling extensions for production deployments.

3-3- Infrastructure Considerations (where to place?)

Cloud architecture plays a pivotal role in VM placement decisions. It encompasses the set of interconnected components and deployment models that define how compute, storage, and network services are delivered. A network-aware placement algorithm must adapt to the physical and logical characteristics of the underlying architecture.

A- Cloud Infrastructure type

Centralized Cloud: infrastructure consolidates all resources in a single data center, offering uniform latency and centralized cooling, power, and network control. Here, placement strategies emphasize intra-rack traffic minimization, server consolidation, and ALR to reduce switch and server energy. Because of the homogeneous environment, algorithms benefit from predictable latencies

Distributed Cloud: infrastructures span multiple, geographically dispersed sites or edge facilities. Placement algorithms in this context must account for WAN latency, variable carbon intensity, renewable energy availability, and differing PUE scores across locations. For instance, placement might favor a solar-powered region despite slightly higher latency. Network-aware algorithms in distributed contexts must balance performance against operational costs and inter-site bandwidth constraints [27]. While distributed placement can optimize global cost and sustainability, it introduces complexity in synchronizing

state across sites, handling network failures, and meeting latency-sensitive SLAs.

B- Cloud Proximity Models

Cloud Proximity Models distinguish between edge and core clouds based on their user-nearness and resource richness.

Edge Clouds: Deployed close to users for latency-sensitive workloads like gaming or AR/VR; placement here must prioritize minimal hop counts and rapid elasticity but suffers from limited capacity and heterogeneous infrastructure. TRACTOR [50], Traffic-aware and Power-efficient Placement in Edge-Cloud Data Centers (ECDCs), an Artificial Bee Colony-based multi-objective VM placement scheme that minimizes network traffic and power consumption in ECDCs. Evaluations on VL2 and three-tier topologies demonstrate a 3.5% reduction in server energy and up to 30% cut in network power usage without degrading QoS. However, TRACTOR presumes accurate pre- and post-placement traffic matrices and requires simulation-based calibration, limiting its adaptability to heterogeneous, real-world edge deployments.

Core Clouds: located in centralized, resource-rich facilities, are suited for compute-heavy, batch-oriented tasks that do not have stringent latency demands. Placement algorithms in these environments optimize resource density and power utilization while managing rack-level heat and congestion. In a centralized high-density core clouds, [51] framework employs a Greedy Randomized VMP (GRVMP) algorithm that fuses heuristic sorting with stochastic perturbations to escape local optima, achieving up to 12% energy reduction and 8% resource utilization gains compared to deterministic baselines. GRVMP addresses dynamic VM arrivals; however, its randomized nature can lead to variability in outcomes and overlooks network topology unless network-aware metrics are integrated.

C- Hardware-Based Energy Mechanisms

Datacenter hardware often embeds energy-saving features at component and network levels. Placement algorithms that are aware of these mechanisms can reduce overall power draw by tailoring VM assignments to exploit them.
We categorize three primary hardware-based strategies below:

- ALR:

ALR dynamically scales the data-link speed of network interfaces (e.g., from 1 Gbps to 100 Mbps) based on instantaneous utilization. When traffic is low, links down-shift to a lower rate—saving up to 40 % of PHY-layer power—then ramp up again under load. Some VM placement schemes explicitly cluster bursty or low-throughput VMs under the same Top-of-Rack switch to maximize low-speed intervals and link-power savings [52].

- DVFS:

Modern CPUs and NICs support DVFS, which lowers voltage and clock frequency when workload demands permit. Experimental studies report up to 30 % server-level energy reduction with minimal performance loss under controlled load variations [53]. Energy-aware schedulers simulate or predict CPU utilization to trigger DVFS states—placing latency-insensitive VMs on hosts where cores can be down-clocked, while reserving full-speed nodes for critical workloads [54].

- Switch and Rack Power-Down:

Many top-of-rack (ToR) switches and rack PDUs can enter sleep modes or shut down unused ports when idle. Research prototypes have shown up to 50 % energy savings in underutilized racks by consolidating traffic and powering down dormant switches [55]. Topology-aware schemes fold traffic into active racks during off-peak periods, allowing idle switches or PDUs to sleep or power off; the migration cost is balanced against the long-term energy gains [56].

Placement algorithms treat ALR, DVFS, and switch/rack power-down not as standalone placement steps but as hardware-aware objectives or constraints that guide where and when to place or migrate VMs. In other words, these features aren't separate "phases" of VM placement; rather, placement algorithms incorporate knowledge of link-rate scaling, voltage/frequency capabilities, or switch on/off thresholds to shape consolidation decisions.
Integrating these hardware-based mechanisms into placement and migration heuristics unlocks significant energy savings that complement software techniques.

D- Thermal-Aware Placement Strategies

Integrating thermal dynamics into VM placement helps prevent hotspots and reduces cooling energy consumption by considering rack- and node-level temperature distributions during allocation and migration decisions [57]. Multi-objective formulations jointly optimize computing energy and cooling load, enabling VM placement algorithms to trade off consolidation benefits against the risk of creating thermal hotspots [58].

3-4- Workload Characteristics (What is being placed)

A- Arrival rate

Static: Static workloads such as batch jobs in scientific computing, benefit from heavy-weight optimizations like ILP, yielding near-optimal resource packing when demands are known in advance [59][60]. The term "static allocation" usually refers to the initial VM placement which is the allocation of VMs to PMs is done during deployment and remains fixed throughout the VMs' lifecycle. The goal is to optimize allocation based on resource requirements and

constraints. However, the assumption of stable load profiles renders it brittle when workloads fluctuate unpredictably.

Dynamic: Dynamic scenarios characterized by real-time VM arrivals in auto-scaling web services or event-driven microservices. Dynamic VM placement includes placing new VMs and migrating existing ones, considering future live migrations, and needs more resources than static solutions.

In this context, reactive placement adapts the initial allocation of resources based on the current state of the system, while proactive placement predicts future conditions and adjusts allocations before problems occur.

- Reactive Placement: Migration or reallocation is triggered by observed thresholds, such as CPU/memory utilization exceeding a limit, network congestion detected on a link, or thermal hot spots. Reactive methods respond to current system state ([61][62]) but often react too late to avoid SLA violations or suboptimal energy states.

- Proactive Placement: Predictive models anticipate future workloads or traffic spikes and migrate VMs preemptively. While more complex, requiring accurate demand prediction, proactive approaches can better prevent overloads and exploit low-utilization windows for consolidation [20], [21]).

B- Workload Type (Application-Centric)

We present the main application categories in the literature used to guide placement heuristics.

Bag of Tasks: Independent parallel tasks requiring minimal inter-communication. Placement focuses on maximizing throughput and minimizing makespan by grouping tasks (VMs) on minimal PMs [41].

CPU-Intensive Workloads: Require sustained processor capacity and thermal stability. Placement must dedicate cores to each VM and move workloads off busy hosts to prevent contention and overheating [64].

Data-Intensive Workloads: Require high I/O and low-latency access to shared storage. Placement must reduce traffic to storage nodes (SNs) and minimize bottlenecks [65].

Latency-Sensitive Applications: Include gaming, financial systems, or telemedicine, where delays severely degrade user experience. These demand edge-aware, low-hop-count VM placement [66].

C- Workload Data Sources for Algorithm Evaluation

The following are the ways researchers evaluate their work against other algorithms. However, researchers may combine two or more types of workload data.

- Benchmark Datasets*:* Standardized collections of VM workload traces detailing CPU, memory, I/O and network usage, collected via monitoring tools, application profilers or user logs. They enable controlled, repeatable comparisons of placement algorithms by quantifying impacts on network utilization, availability and cost.

- Synthetic data: Synthetic data is generated using mathematical models and statistical techniques that simulate the behavior of real-world applications and infrastructure components. It allows researchers to control the workload and resource utilization characteristics of the cloud infrastructure and to compare different algorithms under the same conditions. Researchers evaluated their work using synthetic scenarios with several performance metrics [67].

- Real Traces: Real traces are collected from real cloud computing environments (Amazon EC2, PlanetLab, and Google Cluster) to evaluate VM placement algorithms under realistic conditions. In [51], Amazon EC2 data was used to optimize power consumption. In [68], PlanetLab network traces were utilized to assess algorithm performance. Both methods provide insights into workload behaviors and resource utilization for algorithm evaluation.

These classifications create a multidimensional lens to evaluate VM placement strategies and pave the way for our specialized network-aware taxonomy in Section 4.

4- Taxonomy of Network-aware VM Placement Approaches

This section synthesizes the contextual shifts and motivates the need for a new taxonomy—one that maps VM placement methods not only to their algorithmic families (heuristic, ML-based) but also to the underlying network dynamics they aim to optimize. As shown in Fig.2, our taxonomy therefore introduces a cross-layer perspective that bridges DCN topology, traffic characteristics, communication patterns, and energy reduction strategies, reflecting how emerging solutions should be evaluated in modern cloud environments. Additionally, a sub-taxonomy at the bottom of Fig.2 classifies network-aware VMP algorithms according to their energy consumption strategies.

In a typical cloud computing environment, VMs are interconnected with physical hosts through a network, generating substantial network traffic from the applications they run. Consequently, the placement of VMs on physical hosts significantly impacts network performance, which in turn affects overall application performance. Given that the network is a major consumer of energy, minimizing network traffic and optimizing topology can lead to substantial energy savings.

Therefore, it is critical to consider network-related factors throughout placing and migrating VMs. This means that the VMP algorithm should not only consider the usual metrics

and resource requirements of VMs and PMs but also incorporate network considerations. The algorithm can make more informed decisions regarding VM placement and consolidation by incorporating network conditions, topology, and traffic patterns.

Customers utilize VMs to conduct specific jobs that are frequently parts of larger applications, such as tiers of multi-tier applications. As these VMs start communicating with each other, it can involve the transfer of significant amounts of data, which might increase latency or response times to

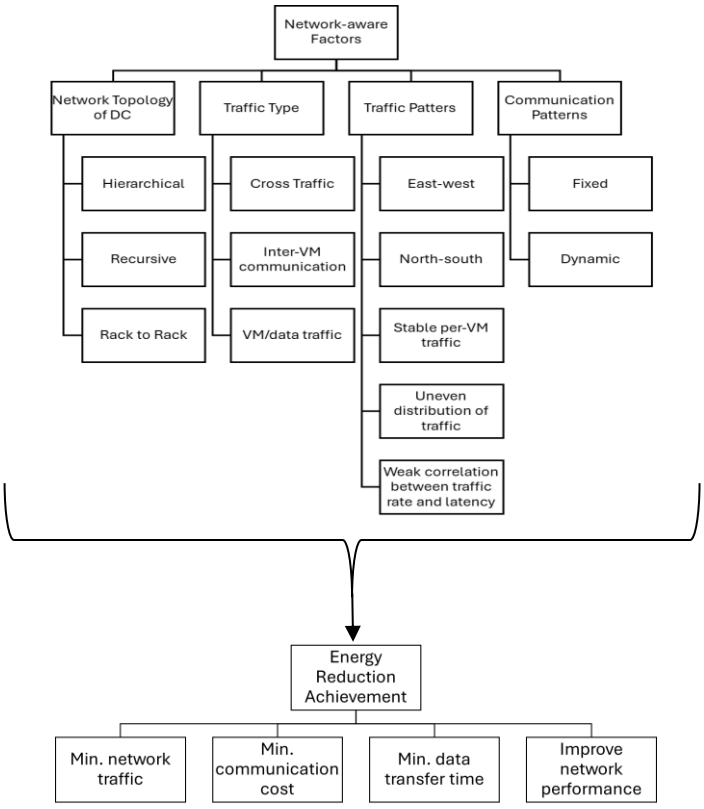

Fig.2 Network-aware VM placement taxonomy.

intolerable levels. In addition, the power consumption of the hardware components involved, such as PMs, routers, switches, and other networking equipment, can also be affected by such communication patterns.

For the reasons listed above, it is ideal to have VMs that communicate frequently placed on the same server, or at the very least within the same DC. Additionally, VMs belonging to the same application may have load correlation, making it more likely that they may peak at the same time; this must also be carefully considered when allocating VM resources.

Network bandwidth can often become a bottleneck, particularly in scenarios involving data mapping on SNs. High network traffic between VMs and SNs can arise when workloads require extensive data mapping. To prevent too many high network loads, it is necessary to consider both the placement of VMs on PMs and application data on SNs. To facilitate this, we categorize network-aware VMP algorithms into four groups based on their focus on network considerations:

4-1- DCN Topology

DC topology involves organizing physical and logical components in a network, including servers, network devices, and SNs. It enables efficient connections with multiple PMs, enhancing energy efficiency and reducing reliability concerns. Various network topologies tackle scalability and energy consumption differently and offer insights for future VM placement research. Researchers can examine the advantages, drawbacks, and enhancements of these topologies to improve current VM placement methods, as discussed in Section V.

A- Hierarchical three-tier

This architecture manages traffic using a structured approach. The access layer connects servers to edge switches, which then relay information to interconnected aggregate switches. The core layer serves as the spine, linking all aggregate switches and handling external connections, providing a scalable and efficient solution for internal data center communication.

- Fat-tree: A three-tier architecture utilizing bipartite graphs with pods as the basic unit, where each pod contains access and aggregation switches. This topology offers efficient routing paths for reducing congestion and power consumption [69].

- VL2: Like fat-tree, this three-tier topology connects core and aggregation switches in a bipartite graph. Valiant load balancing routes traffic by randomly selecting a core switch, reducing congestion and power consumption. A customized VMP technique further optimizes network traffic. [67].

- Portland: This architecture comprises pods with access and aggregation switches forming bipartite graphs, connecting to all core switches. VM placement algorithms prioritize proximity to enhance quality of service (QoS)[70].

B- Recursive

These topologies are constructed recursively, combining smaller building blocks into larger network structures, allowing for scalable and modular designs.

- DCell**:** a server-centric data center network design with a hierarchical structure. Servers connect directly with multiple NICs, organized into cells like cell0, cell1, and cell2 [71].

- BCube**:** BCube is a multi-level data center network architecture focused on servers, integrating them into the network infrastructure. It is derived from hypercube architecture, connecting hosts via switches based on port availability for efficient packet forwarding [72].

C- Rack to Rack

Rack-to-rack networks prioritize communication between server racks. Their design focuses on efficient data transfer within and across racks.

- Scafida**:** a method inspired by scale-free networks to create asymmetric data center topologies with high fault tolerance and small diameters. It allows for flexible scaling but faces challenges with link correlation as the network grows [73].
- Jellyfish: Jellyfish network with random graph topology offers cost-efficiency, 25% more server support, scalability, and flexibility for high-capacity interconnectivity [74].

4-2- Traffic Type

Traffic type categorization in cloud DCs (considered in VMP) optimizes network performance and energy usage by placing VMs with similar traffic types together, reducing data transfers across the network and minimizing energy consumption.

A- Cross-traffic

Cross-traffic is the data flow between VMs or applications that may be located on different servers within the same rack or across different racks. This type of traffic can impact network performance and energy usage. Allocating VMs and data on physically closer PMs can improve efficiency, as explored in [75].

B- Inter-VM communication

North-south traffic involves data flow between virtual machines (VMs) and the Internet, while inter-VM communication refers to data exchange within the same data center. The latter is often high-bandwidth and low-latency, with different application requirements.

Studies are focusing on reducing network energy usage by optimizing VM placement to minimize inter-rack traffic and reduce delays, consequently cutting down on power consumption and costs [76].

C- Traffic between VM and data

This traffic occurs when VMs access data stored on storage devices. VMs send requests to these devices via the network, and the data is transmitted back to the requesting VM. Factors influencing traffic volume include data size, access frequency, and the type of storage device used. In distributed object storage systems, each storage node manages a group of servers. When a server and its corresponding storage node are within the same group, data transfer is optimized, thereby reducing overall traffic flow [77].

4-3- Traffic Patterns

Understanding traffic patterns in cloud networks is crucial for optimizing performance by placing virtual machines in strategic locations to improve network performance and reduce energy consumption. Research indicates that network status changes over time due to unpredictable traffic characteristics, regardless of data center size or type. Authors advocate for traffic-aware VM placement to enhance network scalability by aligning traffic patterns with communication distances. Empirical studies reveal imbalanced communication patterns, link losses, and ON-OFF traffic patterns with varying distributions, emphasizing the need for optimized VM allocation and routing in cloud networks [3] [78].

4-4- Communication Patterns

Communication patterns in VM placement refer to how VMs interact with each other and with external networks. It is a useful resource for perceiving the parallel application communication behavior and is extracted from communication trace, where machines form multiple groups or tiers each of which serves a specific part needed for the accomplishment of the overall task. Energy consumption heavily depends on the communication pattern [79].

A- Fixed

Fixed communication patterns between virtual machines (VMs) exhibit predictable and consistent interactions that remain unchanged during runtime. VM placement strategies often aim to co-locate VMs with frequent communication to minimize network latency and overhead [76].

B- Dynamic

Dynamic communication patterns between VMs change during runtime, in contrast to fixed patterns. This requires adaptable VM placement solutions that monitor and adjust VM locations based on evolving communication needs. The technique introduced in [80] uses a decentralized migration approach considering VM affinity. It dynamically adjusts VM placement through a distributed bartering algorithm to minimize communication overhead and adapt to changing patterns, while maintaining low overhead.

4-5- Energy Reduction Achievement

The energy reduction classification in our taxonomy in Fig.2 is centered around strategies and methodologies in reducing energy consumption in network-aware VM placement. This section highlights how researchers have leveraged network awareness to achieve considerable energy savings in CDCs. In this section, we review different approaches for network traffic minimization, communication cost minimization, data transfer time reduction, and network performance improvement.

A- Minimizing network traffic

One of the effective strategies is to optimize VM placement with the co-location of VMs that communicate with each other with high volume on the same physical hosts. In this way, the distance that data needs to travel is minimal and reduces traffic in the network. For example, the work in [50] suggested a multi-objective VM placement algorithm using a bee colony method, achieving 3.5% power reduction, 15% less network traffic, and 30% lower network power. Similarly, the work in [22] proposed an ant colony optimization algorithm considering both energy usage and network bandwidth, which effectively reduced traffic and outperformed other heuristics.

B- Minimize communication cost

Network communication costs refer to expenses in terms of bandwidth utilization, latency, and rate of data transfer. For VM placement, reducing such costs minimizes resource consumption and overall expenses. The work in [59] introduced a "network consumption" metric to identify optimal VM placements within a fat-tree architecture to minimize network traffic. This approach led to a significant reduction in overall network usage and power consumption, decreasing resource wastage by up to 20%. Similarly, the approach in [81] focused on enhancing VM-to-VM communication using dynamic clustering of VMs based on the network. An adaptive algorithm consolidated VMs to minimize communication costs, leading to reduced high-latency jobs and improved traffic patterns across the network. The goal of these techniques is to strategically place and manage VMs to lower the overall communication costs in the data center network [36].

C- Minimizing Data transfer time

Data transfer time is the duration for data to be transmitted between VMs over the network. It affects energy usage and application performance. Placing VMs closer and grouping them based on traffic patterns can minimize data transfer time. [82] proposed a novel VMP technique that simultaneously improves both VM locations and data rates. They developed heuristics that allocate VMs to PMs with better network bandwidth to reduce the latencies associated with data access. Through simulation experiments, they demonstrated how the proposed approach may lower VMs' data transmission delays.

D- Improving network performance

Improving network performance is the act of optimizing a computer network to enhance its speed, reliability, and efficiency. This involves improving the various components of the network, including switches, routers, cables, servers, and applications, to ensure that data is transmitted quickly, accurately, and consistently. The previously mentioned work in [59] was categorized under minimizing communication cost, but it focused also on minimizing resource wastage, which led to the optimization of the overall network performance.

E- Emerging trends

With the rise of such technologies as network virtualization and Software-Defined Networking (SDN), the way VM placement for energy efficiency will be significantly impacted. Network virtualization increases the flexibility of network resource allocation and management, such that even real-time adjustments according to changing traffic patterns become possible. On the other hand, SDN brings central control to a network, which makes routing much more efficient and leads to lower energy consumption. These technologies are still evolving, we can expect further improvements in energy efficiency and overall network performance in the placement of VMs [83].

5- Discussion

This section discusses the important relationship between network topology, traffic patterns, and energy efficiency in network-aware VMP. We provide a novel perspective on how these aspects interact and affect the total energy consumption within the datacenter.

5-1- Traffic type

Different traffic types have varying requirements regarding reliability, latency, and network bandwidth. For example, real-time communication applications, including video conferencing and VoIP, require low latency and high reliability; in contrast, batch processing applications such as data analytics can tolerate high latency and low reliability. Those network traffic patterns found in datacenters can significantly affect energy consumption, SLAs, cloud provider revenue, as well as the overall cloud infrastructure's efficiency.

In response to such challenges, there has been a development of network-aware VM placement algorithms to optimize network traffic and minimize resource utilization in CDCs. These algorithms distribute the network traffic evenly across the infrastructure to prevent congestion, resulting in energy savings. VMs often rely on the network for data-intensive applications and interactions with other VMs. These algorithms can prioritize high-bandwidth VMs and place them nearby by optimizing the placement of VMs based on their communication patterns, reducing the overall network traffic between and within the data centers. This, in turn, minimizes the number of physical networking components required and leads to reduced power consumption.

5-2- Network topology

Network topology is a principal issue in virtual machine placement, which affects resource utilization and energy efficiency. Placing VMs wisely reduces the distance of data transfers, switches, and links involved in communication and leads to saving energy as well as increasing performance. Fat-tree topology manages the high-bandwidth, low-latency traffic well within a pod or data center, while VL2 is good for traffic generated by VMs in cloud environments, including storage, migration, and inter-DC. BCube is suitable for data-intensive applications that demand high bandwidth and efficient data transmission.

In this subsection, network topology influence on VM placement is discussed based on existing research that examines the impact on energy efficiency as well as overall system performance [84]. The placement of VMs close to each other is quite essential for resource utilization and energy efficiency. Strategic placement reduces the distance of data transfer, therefore reducing the number of switches and links, which means less energy consumption and improved performance in data centers. The three-tier architecture typically includes expensive and power-intensive network devices at the corporate level, whereas DCell and BCube architecture consume similar energy for small-sized data centers. However, BCube consumes more energy for larger data centers. The Fat-Tree topology has reasonable power usage, while BCube is power-intensive due to its extensive use of switches. DCell utilizes commodity switches that consume less power. BCube's design with intermediate servers for routing can pose challenges to energy efficiency.

According to experimental findings, the tree topology experiences congestion issues with similar VM traffic, while the Fat-Tree topology distributes traffic more evenly due to its multi-path connections. VL2 suffers from uneven traffic distribution due to a large gap in link utilization. The Tree topology has lower energy efficiency compared to VL2 and Fat-Tree, although topology awareness can optimize energy usage in the network. However, these conclusions are specific to each author's work, and more research is needed to establish correlations between data center size, server count, switches, and user demands. Cloud service providers should ensure appropriately sized environments to minimize costs. A hybrid or dynamic topology approach using SDN can optimize resource utilization, energy efficiency, and overall performance by adapting the network topology based on workload demands, such as favoring a fat-tree topology for high east-west traffic.

5-3- Traffic and Communication patterns

To minimize energy consumption in DCs, network-aware VM placement algorithms play a crucial role. These algorithms aim to allocate VMs with similar traffic patterns to the same physical servers or switches. This will reduce inter-server or inter-switch communication, therefore saving energy not only in the network infrastructure but also in the servers. Secondly, VMP optimization based on bandwidth and latency demands will prevent network congestion, thus assuring satisfactory performance and energy efficiency during communications.

Energy consumption and network traffic in virtualized environments were analyzed in studies [58,59]. It was noticed that energy consumption might have a wide variation for different traffic allocation strategies and that the type of traffic may strongly influence the possible energy savings. Such results are important to consider in traffic-aware optimizations, but all such optimizations require detailed information from clients about the application network and communication requirements. This allows network-aware techniques for minimizing communication delays and/or improving overall application performance.

The distribution of the components over various PMs provides a good opportunity for parallel processing in applications such as MapReduce. In case migration needs to be done, the ideal order of the intercommunicating virtual machines will help avoid core network traffic and energy consumption. Considering intercommunication between replicated virtual machines is also important to prevent bottlenecks and excessive energy usage.

Recognition of the traffic pattern is especially important in dynamic cloud environments. Workload and communication requirements are dynamic; hence, the adaptability of VMP algorithms is required to achieve resource and energy efficiency. Such dynamical traffic management approaches like load balancing and traffic shaping would prevent congestion and optimize power consumption.

The application-specific information will also reduce latency, inter-VM traffic, and improve application performance in placement algorithms. On the other hand, machine learning algorithms will use historical traffic data and predictive models to foresee traffic patterns, thus making proactive placement decisions that reduce energy consumption. Machine learning can also help in identifying and classifying traffic hotspots, which helps in applying targeted optimizations to mitigate power imbalances.

6- Conclusion And Future Directions

This paper presents a new classification for VM placement techniques in CDCs that are both network-aware and energy-efficient. It examines various network factors, including network equipment, workload type, performance, scalability, efficiency, reliability, and availability, to understand how VM placement affects network performance. The research indicates that network-aware VM placement algorithms can boost performance by reducing latency between VMs and improving security through co-location. However, the initial deployment of these algorithms might incur higher costs, necessitating a careful evaluation of the trade-off between energy consumption and migration costs.

This work also reviews research that identifies the most effective metrics for evaluating the performance of network-aware VM placement algorithms, focusing on energy efficiency, network performance, and resource utilization. Additionally, the study examines how network topology affects energy consumption in data centers and the trade-off between energy use and migration costs, providing valuable insights. These insights can help researchers develop and implement more effective network-aware VM placement algorithms that optimize energy consumption, improve network performance, and minimize migration costs. Based on the findings, future research directions for network-aware VM placement in CDCs can be suggested, including:

- Developing energy-efficient algorithms that consider the network metrics identified in this study. This would involve creating strategies to optimize energy use while improving network performance, factoring in elements like datacenter layout and communication patterns.

- Testing VM placement techniques on realistic testbeds. While simulations help assess the proposed VM placement methods, it is essential to validate these techniques on actual cloud testbeds with real-world network topologies.

- Researching VM placement algorithms that enhance security and privacy in cloud environments. This could involve devising methods to group related VMs on the same server or rack while preventing the co-location of unrelated VMs. Such strategies would help mitigate the risk of security breaches and protect sensitive data in cloud settings.

- Continuing to explore novel solutions for optimizing VM placement and migration that can boost energy efficiency and network performance in CDCs. This would include investigating innovative techniques and approaches that leverage emerging technologies like machine learning and artificial intelligence to improve network-aware VM placement.

Future research in this area could investigate how elements like energy storage systems, renewable energy sources, and workload balancing impact network-aware VM placement. These potential directions provide a solid foundation for further exploration of energy-efficient network-aware VM placement, intending to create more effective strategies for optimizing energy consumption, improving network performance, enhancing security and privacy, and integrating artificial intelligence throughout the cloud computing environment.

References

[1] P. M. Mell and T. Grance, “The NIST definition of cloud computing,” Gaithersburg, MD, 2011. doi: 10.6028/NIST.SP.800-145.

[2] D. Bliedy, S. Mazen, and E. Ezzat, “Datacentre Total Cost of Ownership (TCO) Models : A Survey,” International Journal of Computer Science, Engineering and Applications, vol. 8, no. 2/3/4, pp. 47–62, 2018, doi: 10.5121/ijcsea.2018.8404.

[3] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, pp. 267–280, 2010, doi: 10.1145/1879141.1879175.

[4] L. Zhou, C. H. Chou, L. N. Bhuyan, K. K. Ramakrishnan, and D. Wong, “Joint server and network energy saving in data centers for latency-sensitive applications,” Proceedings - 2018 IEEE 32nd International Parallel and Distributed Processing Symposium, IPDPS 2018, pp. 700–709, 2018, doi: 10.1109/IPDPS.2018.00079.

[5] K. Bilal et al., “A survey on Green communications using Adaptive Link Rate,” Cluster Comput, vol. 16, no. 3, pp. 575–589, Jul. 2013, doi: 10.1007/s10586-012-0225-8.

[6] A. C. Orgerie, M. D. De Assuncao, and L. Lefevre, “A survey on techniques for improving the energy efficiency of large-scale distributed systems,” ACM Comput Surv, vol. 46, no. 4, 2014, doi: 10.1145/2532637.

[7] M. H. Ferdaus, M. Murshed, R. N. Calheiros, and R. Buyya, Network-aware virtual machine placement and migration in cloud data centers, no. May. 2015. doi: 10.4018/978-1-4666-8213-9.ch002.

[8] A. Hammadi and L. Mhamdi, “A survey on architectures and energy efficiency in Data Center Networks,” Comput Commun, vol. 40, pp. 1–21, 2014, doi: 10.1016/j.comcom.2013.11.005.

[9] K. Bilal et al., “A taxonomy and survey on Green Data Center Networks,” Future Generation Computer Systems, vol. 36, pp. 189–208, Jul. 2014, doi: 10.1016/j.future.2013.07.006.

[10] R. W. Ahmad, A. Gani, S. H. A. Hamid, M. Shiraz, A. Yousafzai, and F. Xia, “A survey on virtual machine migration and server consolidation frameworks for cloud data centers,” Journal of Network and Computer Applications, vol. 52, pp. 11–25, 2015, doi: 10.1016/j.jnca.2015.02.002.

[11] F. L. Pires and B. Baran, “A virtual machine placement taxonomy,” Proceedings - 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2015, no. July, pp. 159–168, 2015, doi: 10.1109/CCGrid.2015.15.

[12] M. Masdari, S. S. Nabavi, and V. Ahmadi, “An overview of virtual machine placement schemes in cloud computing,” Journal of Network and Computer Applications, vol. 66, pp. 106–127, 2016, doi: 10.1016/j.jnca.2016.01.011.

[13] H. Talebian et al., Optimizing virtual machine placement in IaaS data centers: taxonomy, review and open issues, vol. 23, no. 2. Springer US, 2020. doi: 10.1007/s10586-019-02954-w.

[14] H. Zhuang and B. Esmaeilpour Ghouchani, "Virtual machine placement mechanisms in the cloud environments: a systematic review," Kybernetes, vol. 50, no. 2, pp. 333–368, 2021, doi: 10.1108/K-09-2019-0635.

[15] L. Helali and M. N. Omri, "A survey of data center consolidation in cloud computing systems," 2021. doi: 10.1016/j.cosrev.2021.100366.

[16] A. Sumathi, … B. K.-T. J. of, and undefined 2023, "Advancements in Energy-Efficient Virtual Machine Placement Survey for Cloud Computing," Researchgate.Net, no. February, 2024, doi: 10.13140/RG.2.2.17918.36164.

[17] N. Rana et al., "A systematic literature review on contemporary and future trends in virtual machine scheduling techniques in cloud and multi-access computing," Front Comput Sci, vol. 6, 2024, doi: 10.3389/fcomp.2024.1288552.

[18] J. Zou, K. Wang, K. Zhang, and M. Kassim, "Perspective of virtual machine consolidation in cloud computing: a systematic survey," Telecommun Syst, p. 11235, 2024, doi: 10.1007/s11235-024-01184-9.

[19] S. R. Swain, A. Parashar, A. K. Singh, and C. Nan Lee, "An Energy Efficient Virtual Machine Placement Scheme for Intelligent Resource Management at Cloud Data Center," in OCIT 2023 - 21st International Conference on Information Technology, Proceedings, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 65–70. doi: 10.1109/OCIT59427.2023.10430915.

[20] S. Kumar, S. Mittal, and M. Singh, "Active VM Placement Approach Based on Energy Efficiency in Cloud Environment," in Lecture Notes in Networks and Systems, Springer Science and Business Media Deutschland GmbH, 2022, pp. 35–46. doi: 10.1007/978-981-19-1018-0_4.

[21] Z. Li, K. Lin, S. Cheng, L. Yu, and J. Qian, "Energy-Efficient and Load-Aware VM Placement in Cloud Data Centers," J Grid Comput, vol. 20, no. 4, 2022, doi: 10.1007/s10723-022-09631-0.

[22] H. Xing, J. Zhu, R. Qu, P. Dai, S. Luo, and M. A. Iqbal, "An ACO for energy-efficient and traffic-aware virtual machine placement in cloud computing," Swarm Evol Comput, vol. 68, no. November 2021, p. 101012, 2022, doi: 10.1016/j.swevo.2021.101012.

[23] D. Dabhi and D. Thakor, "Utilisation-aware VM placement policy for workload consolidation in cloud data centres," International Journal of Communication Networks and Distributed Systems, vol. 28, no. 6, pp. 704–726, 2022, doi: 10.1504/ijcnds.2022.126224.

[24] E. I. Elsedimy, M. Herajy, and S. M. M. Abohashish, "Energy and QoS-aware virtual machine placement approach for IaaS cloud datacenter," 2025. doi: 10.1007/s00521-024-10872-1.

[25] K. Lu, R. Yahyapour, P. Wieder, C. Kotsokalis, E. Yaqub, and A. I. Jehangiri, "QoS-aware VM placement in multi-domain service level agreements scenarios," IEEE International Conference on Cloud Computing, CLOUD, no. April 2014, pp. 661–668, 2013, doi: 10.1109/CLOUD.2013.112.

[26] T. Renugadevi, K. Geetha, K. Muthukumar, and Z. W. Geem, "Optimized energy cost and carbon emission-aware virtual machine allocation in sustainable data centers," Sustainability (Switzerland), vol. 12, no. 16, pp. 1–27, 2020, doi: 10.3390/SU12166383.

[27] S. Rawas, A. Zekri, and A. El Zaart, "Power and Cost-Aware Virtual Machine Placement in Geo-Distributed Data Power and Cost-aware Virtual Machine Placement in Geo-distributed Data Centers," no. March, 2018, doi: 10.5220/0006696201120123.

[28] G. P. Maskare and S. Sharma, "The Hybrid ACO, PSO, and ABC Approach for Load Balancing in Cloud Computing," vol. 10, 2023, Accessed: May 07, 2025. [Online]. Available: www.jetir.org

[29] M. H. Kim, J. Y. Lee, S. A. Raza Shah, T. H. Kim, and S. Y. Noh, "Min-max exclusive virtual machine placement in cloud computing for scientific data environment," Journal of Cloud Computing, vol. 10, no. 1, pp. 1–17, Dec. 2021, doi: 10.1186/S13677-020-00221-7/FIGURES/12.

[30] M. Koubàa, R. Regaieg, A. S. Karar, M. Nadeem, and F. Bahloul, "A Multi-Objective Approach for Optimizing Virtual Machine Placement Using ILP and Tabu Search," Telecom, vol. 5, no. 4, pp. 1309–1331, 2024, doi: 10.3390/telecom5040065.

[31] X. Zheng and Y. Xia, "Exploring mixed integer programming reformulations for virtual machine placement with disk anti-colocation constraints," Performance Evaluation, vol. 135, 2019, doi: 10.1016/j.peva.2019.102035.

[32] S. Yang, P. Wieder, R. Yahyapour, S. Trajanovski, and X. Fu, "Reliable Virtual Machine Placement and Routing in Clouds," IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 10, pp. 2965–2978, 2017, doi: 10.1109/TPDS.2017.2693273.

[33] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing," Future Generation Computer Systems, vol. 28, no. 5, pp. 755–768, 2012, doi: 10.1016/j.future.2011.04.017.

[34] J. Wang, J. Yu, R. Zhai, X. He, and Y. Song, "GMPR: A Two-Phase Heuristic Algorithm for Virtual Machine Placement in Large-Scale Cloud Data Centers," IEEE Syst J, vol. 17, no. 1, pp. 1419–1430, Mar. 2023, doi: 10.1109/JSYST.2022.3187971.

[35] S. Jangiti, V. Vijayakumar, and V. Subramaniyaswamy, "Hybrid best-fit heuristic for energy efficient virtual machine placement in cloud data centers," EAI Endorsed Transactions on Energy Web, vol. 7, no. 26, pp. 1–7, 2020, doi: 10.4108/eai.13-7-2018.162689.

[36] R. Keshri and D. P. Vidyarthi, "Communication-aware, energy-efficient VM placement in cloud data center using ant colony optimization," International Journal of Information Technology (Singapore), vol. 15, no. 8, pp. 4529–4535, Dec. 2023, doi: 10.1007/S41870-023-01531-0/METRICS.

[37] N. Donyagard Vahed, M. Ghobaei-Arani, and A. Souri, "Multiobjective virtual machine placement mechanisms using nature-inspired metaheuristic algorithms in cloud environments: A comprehensive review," International Journal of Communication Systems, vol. 32, no. 14, 2019, doi: 10.1002/dac.4068.

[38] A. S. Abohamama and E. Hamouda, "A hybrid energy–Aware virtual machine placement algorithm for cloud environments," Expert Syst Appl, vol. 150, p. 113306, 2020, doi: 10.1016/j.eswa.2020.113306.

[39] A. M. Baydoun and A. S. Zekri, "Network-, Cost-, and Renewable-Aware Ant Colony Optimization for Energy-Efficient Virtual Machine Placement in Cloud Datacenters," Future Internet, vol. 17, no. 6, p. 261, Jun. 2025, doi: 10.3390/fi17060261.

[40] S. Talwani et al., "Machine-Learning-Based Approach for Virtual Machine Allocation and Migration," Electronics (Switzerland), vol. 11, no. 19, 2022, doi: 10.3390/electronics11193249.

[41] S. Rawas, A. Zekri, and A. El-Zaart, "LECC: Location, energy, carbon and cost-aware VM placement model in geo-distributed DCs," Sustainable Computing: Informatics and Systems, vol. 33, 2022, doi: 10.1016/j.suscom.2021.100649.

[42] A. Jumnal and S. M. Dilip Kumar, "Optimal VM placement approach using fuzzy reinforcement learning for cloud data centers," in Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 29–35. doi: 10.1109/ICICV50876.2021.9388424.

[43] H. Padmanaban, "Machine Learning Algorithms Scaling on Large-Scale Data Infrastructure," Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023, vol. 3, no. 1, pp. 1–26, Apr. 2024, doi: 10.60087/JAIGS.VOL03.ISSUE01.P26.

[44] H. A. Alharbi, T. E. H. Elgorashi, A. Q. Lawey, and J. M. H. Elmirghani, "The Impact of Inter-Virtual Machine Traffic on Energy Efficient Virtual Machines Placement," in 2019 IEEE Sustainability through ICT Summit, StICT 2019, 2019. doi: 10.1109/STICT.2019.8789381.

[45] F. kamoun-abid, H. Frikha, A. Meddeb-Makhoulf, and F. Zarai, "Allocation of virtual machine in a cloud environment based on machine learning," Res Sq, Jan. 2023, doi: 10.21203/RS.3.RS-2483861/V1.

[46] N. Tziritas, T. Loukopoulos, S. Khan, C. Z. Xu, and A. Zomaya, "A communication-aware energy-efficient graph-coloring algorithm for VM placement in clouds," Proceedings - 2018 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovations, SmartWorld/UIC/ATC/ScalCom/CBDCo, pp. 1684–1691, 2018, doi: 10.1109/SmartWorld.2018.00286.

[47] S. Sadegh, K. Zamanifar, P. Kasprzak, and R. Yahyapour, "A two-phase virtual machine placement policy for data-intensive applications in cloud," Journal of Network and Computer Applications, vol. 180, no. December 2020, p. 103025, 2021, doi: 10.1016/j.jnca.2021.103025.

[48] J. Gedeon, M. Stein, L. Wang, and M. Mühlhäuser, "On Scalable In-Network Operator Placement for Edge Computing".

[49] T. Huang, W. Huang, B. Zhang, W. Chen, and X. Pan, "Optimizing energy consumption in centralized and distributed cloud architectures with a comparative study to increase stability and efficiency," Energy Build, vol. 333, 2025, doi: 10.1016/j.enbuild.2025.115454.

[50] S. S. Nabavi, S. S. Gill, M. Xu, M. Masdari, and P. Garraghan, "TRACTOR: Traffic-aware and power-efficient virtual machine placement in edge-cloud data centers using artificial bee colony optimization," International Journal of Communication Systems, vol. 35, no. 1, pp. 1–20, 2022, doi: 10.1002/dac.4747.

[51] S. Azizi, M. Shojafar, J. Abawajy, and R. Buyya, "GRVMP: A Greedy Randomized Algorithm for Virtual Machine Placement in Cloud Data Centers," IEEE Syst J, vol. 15, no. 2, pp. 2571–2582, 2020, doi: 10.1109/jsyst.2020.3002721.

[52] W. Wei, H. Gu, W. Lu, T. Zhou, and X. Liu, "Energy Efficient Virtual Machine Placement with an Improved Ant Colony Optimization over Data Center Networks," IEEE Access, vol. 7, pp. 60617–60625, 2019, doi: 10.1109/ACCESS.2019.2911914.

[53] S. Bani-Ahmad, S. Sa'adeh, S. Bani-Ahmad, and S. Sa'adeh, "Scalability of the DVFS Power Management Technique as Applied to 3-Tier Data Center Architecture in Cloud Computing," Journal of Computer and Communications, vol. 5, no. 1, pp. 69–93, Dec. 2016, doi: 10.4236/JCC.2017.51007.

[54] J. Masoudi, B. Barzegar, and H. Motameni, "Energy-Aware Virtual Machine Allocation in DVFS-Enabled Cloud Data Centers," IEEE Access, vol. 10, pp. 3617–3630, 2022, doi: 10.1109/ACCESS.2021.3136827.

[55] "ElasticTree: Saving Energy in Data Center Networks," in Proceedings of the 7th USENIX Symposium on Networked Systems Design and Implementation, San Jose, CA, USA, Apr. 2010.

[56] S. Xiao, Y. Cui, X. Wang, Z. Yang, S. Yan, and L. Yang, "Traffic-aware Virtual Machine Migration in Topology-adaptive DCN," Proceedings - International Conference on Network Protocols, ICNP, vol. 2016-December, Dec. 2016.

[57] A. Akbari, A. Khonsari, and S. M. Ghoreyshi, "Thermal-aware virtual machine allocation for heterogeneous cloud data centers," Energies (Basel), vol. 13, no. 11, 2020, doi: 10.3390/en13112880.

[58] J. Lin, W. Lin, W. Wu, W. Lin, and K. Li, "Energy-aware virtual machine placement based on a holistic thermal model for cloud data centers," Future Generation Computer Systems, vol. 161, pp. 302–314, 2024, doi: 10.1016/j.future.2024.07.020.

[59] S. Omer, S. Azizi, M. Shojafar, and R. Tafazolli, "A priority, power and traffic-aware virtual machine placement of IoT applications in cloud data centers," Journal of Systems Architecture, vol. 115, no. April, 2021, doi: 10.1016/j.sysarc.2021.101996.

[60] A. K. Singh, S. R. Swain, D. Saxena, and C. N. Lee, "A Bio-Inspired Virtual Machine Placement Toward Sustainable Cloud Resource Management," IEEE Syst J, vol. 17, no. 3, pp. 3894–3905, 2023, doi: 10.1109/JSYST.2023.3248118.

[61] H. F. Farimani, S. R. K. Tabbakh, D. Bahrepour, and R. Ghaemi, "Reallocation of virtual machines to cloud data centers reduce service level agreement violation and energy consumption using the FMT method," Journal of Information Systems and Telecommunication, vol. 7, no. 4, pp. 316–325, 2019.

[62] F. Alharbi, Y. C. Tian, M. Tang, W. Z. Zhang, C. Peng, and M. Fei, "An Ant Colony System for energy-efficient dynamic Virtual Machine Placement in data centers," Expert Syst Appl, vol. 120, pp. 228–238, 2019, doi: 10.1016/j.eswa.2018.11.029.

[63] S. Mashhadi Moghaddam, M. O'Sullivan, C. Walker, S. Fotuhi Piraghaj, and C. P. Unsworth, "Embedding individualized machine learning prediction models for energy efficient VM consolidation within Cloud data centers," Future Generation Computer Systems, vol. 106, pp. 221–233, 2020, doi: 10.1016/j.future.2020.01.008.

[64] A. Kamalinia and A. Ghaffari, "Hybrid Task Scheduling Method for Cloud Computing by Genetic and PSO Algorithms," Journal of Information Systems and

Telecommunication, vol. 4, no. 16, pp. 1–10, 2017, doi: 10.1007/s11277-017-4839-2.

[65] S. Sadegh, K. Zamanifar, P. Kasprzak, and R. Yahyapour, "A two-phase virtual machine placement policy for data-intensive applications in cloud," Journal of Network and Computer Applications, vol. 180, p. 103025, Apr. 2021, doi: 10.1016/J.JNCA.2021.103025.

[66] Y. Fan, H. Ding, L. Wang, and X. Yuan, "Green latency-aware data placement in data centers," Computer Networks, vol. 110, pp. 46–57, 2016, doi: 10.1016/j.comnet.2016.09.015.

[67] S. Farzai, M. H. Shirvani, and M. Rabbani, "Communication-Aware Traffic Stream Optimization for Virtual Machine Placement in Cloud Datacenters with VL2 Topology," no. May, 2021.

[68] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers," Concurrency Computation Practice and Experience, vol. 24, no. 13, pp. 1397–1420, 2012, doi: 10.1002/cpe.1867.

[69] S. Fang, R. Kanagavelu, B. S. Lee, C. H. Foh, and K. M. M. Aung, "Power-efficient virtual machine placement and migration in data centers," Proceedings - 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, GreenCom-iThings-CPSCom 2013, pp. 1408–1413, 2013, doi: 10.1109/GreenCom-iThings-CPSCom.2013.246.

[70] S. Georgiou, K. Tsakalozos, and A. Delis, "Exploiting network-topology awareness for VM placement in IaaS clouds," in Proceedings - 2013 IEEE 3rd International Conference on Cloud and Green Computing, CGC 2013 and 2013 IEEE 3rd International Conference on Social Computing and Its Applications, SCA 2013, 2013, pp. 151–158. doi: 10.1109/CGC.2013.30.

[71] "Data center network architectures." [Online]. Available: https://en.wikipedia.org/wiki/Data_center_network_architectures

[72] C. Guo et al., "BCube: A high performance, server-centric network architecture for modular data centers," Computer Communication Review, vol. 39, no. 4, pp. 63–74, 2009, doi: 10.1145/1594977.1592577.

[73] L. Gyarmati and T. A. Trinh, "Scafida: A scale-free network inspired data center architecture," 2010. doi: 10.1145/1880153.1880155.

[74] A. Singla, C. Y. Hong, L. Popa, and P. B. Godfrey, "Jellyfish: Networking data centers randomly," Proceedings of NSDI 2012: 9th USENIX Symposium on Networked Systems Design and Implementation, pp. 225–238, 2012.

[75] M. C. Çavdar, I. Korpeoglu, and Ö. Ulusoy, "A Utilization Based Genetic Algorithm for virtual machine placement in cloud systems," Comput Commun, vol. 214, pp. 136–148, Jan. 2024, doi: 10.1016/J.COMCOM.2023.11.028.

[76] K. Lacurts, S. Deng, A. Goyal, and H. Balakrishnan, "Choreo: Network-Aware Task Placement for Cloud Applications," 2013, doi: 10.1145/2504730.2504744.

[77] Q. Zheng et al., "Virtual machine consolidated placement based on multi-objective biogeography-based optimization," Future Generation Computer Systems, vol. 54, pp. 95–122, Jan. 2016, doi: 10.1016/J.FUTURE.2015.02.010.

[78] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding Data Center Traffic Characteristics," in Computer Communication Review, 2010, pp. 92–99.

[79] S. M. Nabavinejad and M. Goudarzi, "Communication-Awareness for Energy- Efficiency in Datacenters," in Advances in Computers, vol. 100, 2016, pp. 201–254.

[80] J. Sonnek, J. Greensky, R. Reutiman, and A. Chandra, "Starling: Minimizing Communication Overhead in Virtualized Computing Platforms Using Decentralized Affinity-Aware Migration ," 2009.

[81] G. Luo, Z. Qian, M. Dong, K. Ota, and S. Lu, "Improving performance by network-aware virtual machine clustering and consolidation," Journal of Supercomputing, vol. 74, no. 11, pp. 5846–5864, 2018, doi: 10.1007/s11227-017-2104-9.

[82] K. Zamanifar, N. Nasri, and M. H. Nadimi-Shahraki, "Data-aware virtual machine placement and rate allocation in cloud environment," Proceedings - 2012 2nd International Conference on Advanced Computing and Communication Technologies, ACCT 2012, pp. 357–360, 2012, doi: 10.1109/ACCT.2012.40.

[83] S. Aggarwal, N. Kumar, S. Tanwar, and M. Alazab, "A Survey on Energy Trading in the Smart Grid: Taxonomy, Research Challenges and Solutions," IEEE Access, vol. 9, pp. 116231–116253, 2021, doi: 10.1109/access.2021.3104354.

[84] J. K. Dong, H. B. Wang, Y. Y. Li, and S. D. Cheng, "Virtual machine placement optimizing to improve network performance in cloud data centers," Journal of China Universities of Posts and Telecommunications, vol. 21, no. 3, pp. 62–70, 2014, doi: 10.1016/S1005-8885(14)60302-2.

[85] C. Xu, Z. Zhao, H. Wang, R. Shea, and J. Liu, "Energy Efficiency of Cloud Virtual Machines: From Traffic Pattern and CPU Affinity Perspectives," IEEE Syst J, vol. 11, no. 2, pp. 835–845, 2017, doi: 10.1109/JSYST.2015.2429731.