Shahid Bahonar University of Kerman
Journal of Mahani Mathematical Research
Print ISSN: 2251-7952 Online ISSN: 2645-4505

THERMAL-AWARE VIRTUAL MACHINE PLACEMENT APPROACHES: A SURVEY

S. Dadashi ✉ and A. Aghasi

Article type: Research Article

(Received: 21 August 2023, Received in revised form 18 February 2024)

(Accepted: 28 April 2024, Published Online: 14 May 2024)

Abstract. Thermal-aware virtual machine (VM) placement has emerged as a critically significant research domain in response to the escalating demand for energy-efficient and dependable cloud data centers. Addressing the imperative need for resource optimization and reduced energy consumption, the virtual machine placement problem seeks to strategically allocate VMs to physical servers while adhering to stringent thermal constraints. This paper intricately surveys the state-of-the-art techniques employed in thermal-aware VM placement, encompassing both static and dynamic approaches. Our comprehensive analysis delves into influential factors, including workload characteristics, server heterogeneity, and advanced thermal management techniques. By elucidating the intricacies of these considerations, our review offers a nuanced understanding of the complex VM placement landscape. Importantly, we spotlight key challenges and identify open research issues, presenting a roadmap for future investigations. This review paper stands as a pivotal resource, providing invaluable insights for researchers and practitioners navigating the evolving landscape of thermal-aware virtual machine placement in cloud data centers.

Keywords: cloud computing, energy saving, thermal awareness, virtual machine, data center.
2020 MSC: 68M14, 68M15

1. Introduction

As the use of cloud computing continues to grow, so too does the demand for efficient and cost-effective ways to manage the resources of data centers. One important aspect of this management is virtual machine (VM) placement, which involves determining where to allocate VMs within a data center to maximize resource utilization and minimize energy consumption [1]. In recent years, there has been a growing interest in thermal-aware VM placement, which takes into account the thermal characteristics of the data center when making placement decisions [2]. This approach recognizes that the temperature of a data center can have a significant impact on its overall performance and energy

✉ sajed.dadashi@gmail.com, ORCID: 0000-0002-2192-315X
https://doi.org/10.22103/jmmr.2024.22069.1497
Publisher: Shahid Bahonar University of Kerman
How to cite: S. Dadashi, A. Aghasi, *Thermal-aware virtual machine placement approaches: A survey*, J. Mahani Math. Res. 2024; 13(2): 325 - 363.

consumption [3]. Thermal-aware VM placement involves analyzing various factors such as the cooling system, the layout of the data center, and the thermal profiles of individual VMs. By considering these factors, it becomes possible to make more informed decisions about where to place VMs in order to optimize resource utilization while maintaining appropriate temperature levels [4]. Overall, thermal-aware VM placement represents an important area of research in cloud computing and data center management [5]. By incorporating thermal considerations into VM placement decisions, it is possible to improve the efficiency and sustainability of data centers while ensuring reliable performance for users. In this paper, we aim to provide a comprehensive review of the current state of research in thermal-aware VM placement for cloud computing data centers. We will begin by discussing the importance of thermal management in data centers and the challenges associated with optimizing resource utilization while maintaining appropriate temperature levels [6]. We will then review existing literature on thermal-aware VM placement, including different approaches and techniques that have been designed to address this issue. Next, we will present a detailed analysis of the various factors that must be considered when implementing thermal-aware VM placement, including the layout of the data center and the cooling system [7]. The final goal is to provide a comprehensive overview of this important area of research and to identify new directions for future work in thermal-aware VM placement for cloud computing data centers [8]. The traditional placement procedures can be implemented either statically or dynamically. fundamental differences and the advantages and disadvantages of each approach are as follows:

Static Approach: In the static approach, VM placement decisions are made at the initial deployment stage and remain fixed for an extended period. This approach considers historical data or predefined rules to determine the optimal placement of VMs based on thermal criteria.
Advantages:
- Simplicity: Static approaches are relatively simple to implement as they do not require continuous monitoring or real-time adjustments.
- Low overhead: Once the initial placement is determined, there is minimal computational overhead or runtime costs associated with VM migrations.
Disadvantages:
- Lack of adaptability: Static approaches do not account for changes in workload dynamics, temperature fluctuations, or server failures, which can lead to suboptimal placements over time. - Inefficiency: Fixed placements may result in hotspots or uneven heat distribution as the workload or environmental conditions evolve.

Dynamic Approach:
The dynamic approach involves continuously monitoring the thermal conditions of the data center and dynamically adjusting the VM placement to optimize

thermal management. This approach relies on real-time data, predictive analytics and dynamic placement algorithms to adapt to changing workload and thermal conditions.
Advantages:
- Adaptability: Dynamic approaches can react to real-time temperature variations, workloads, and environmental changes, ensuring optimal VM placement.
- Thermal resilience: Continuous monitoring and adjustment help prevent hotspots and improve overall thermal management. - Improved energy efficiency: Dynamic approaches can leverage workload consolidation and VM migrations to optimize energy consumption for cooling.
Disadvantages:
- Increased complexity: Dynamic approaches require sophisticated monitoring systems, real-time data analytics, and intelligent placement algorithms, leading to increased complexity in implementation.
- Migration overhead: Dynamic VM placement may involve frequent VM migrations, which can introduce network and computational overhead, potentially impacting performance and user experience. While static approaches offer simplicity and low runtime costs, dynamic approaches provide adaptability, thermal resilience, and improved energy efficiency. The choice between the two depends on the specific requirements of the data center and the trade-off between flexibility and complexity. Many researchers are exploring hybrid approaches that combine the advantages of both static and dynamic techniques to achieve more robust thermal-aware VM placement algorithms.

On the other hand, new approaches in thermal-aware VM placement can be divided as follows:

1. Machine learning-based approach: This approach involves using machine learning algorithms to predict workload patterns and thermal behavior of the data center. The algorithm can then optimize virtual machine placement based on these predictions to ensure energy efficiency and thermal management.

2. Hybrid approach: This approach combines static and dynamic techniques to achieve better performance in virtual machine placement. The static approach can be used to optimize the initial placement, while the dynamic approach can be used to adjust the placement based on real-time workload and thermal conditions.

3. Multi-objective optimization approach: This approach considers multiple objectives, such as energy consumption, resource utilization, and thermal constraints, simultaneously. It uses optimization techniques to find the best trade-off between these objectives, leading to an optimal virtual machine placement.

4. Edge computing-based approach: This approach involves placing virtual machines closer to the end-users, reducing the need for data transfer and improving response time. It also considers the thermal constraints of edge devices and optimizes virtual machine placement accordingly.

5. Blockchain-based approach: This approach involves using blockchain technology to enable decentralized virtual machine placement decisions. It ensures transparency, security, and fairness in the decision-making process, leading to an optimal virtual machine placement.

The virtual machine placement process in data centers is influenced by some critical factors such as workload characteristics and server heterogeneity. Workload characteristics encompass the specific requirements and characteristics of applications, such as CPU, memory, I/O patterns, and network traffic. Server heterogeneity refers to the differences in hardware capabilities among servers, including CPU, memory, storage, and network resources. These factors impact the decision-making process for virtual machine placement in several ways. Firstly, they optimize resource allocation by ensuring that each virtual machine is placed on a server with adequate resources to meet its requirements, thus maximizing performance and minimizing resource contention. Secondly, they facilitate load balancing by distributing workloads evenly across servers, leveraging server heterogeneity to prevent overloads or under utilization. Lastly, they enable energy efficiency considerations by consolidating CPU-intensive workloads onto energy-efficient servers, reducing overall power consumption. By considering workload characteristics and server heterogeneity, the placement process strives to optimize resource usage, balance workloads, and improve energy efficiency.

1.1. CONTRIBUTION OF THE PAPER VS. OTHER REVIEW PAPERS

. the contributions of this paper are focused on thermal-aware virtual machine (VM) placement techniques, considering various aspects of the decision-making process, and incorporating the use of Software-Defined Networking (SDN) technology. These contributions enhance our understanding of the factors influencing the VM placement process, highlight challenges and open research issues, and provide future directions for researchers to explore.

1. Comprehensive review of thermal-aware VM placement techniques: this paper provides a comprehensive review of VM placement techniques that consider thermal considerations. This review summarizes existing methodologies, algorithms, and approaches proposed in the literature for thermal-aware placement. By evaluating and analyzing these techniques, your paper offers valuable insights into the state-of-the-art practices for managing temperature-related issues in data centers.

2. Analysis of factors influencing VM placement process: In addition to thermal-aware placement, this paper analyzes the various factors that impact the VM placement process. This analysis takes into account workload characteristics, such as CPU and memory requirements, I/O patterns, and network traffic. Moreover, it explores the influence of server heterogeneity, which encompasses the differences in hardware configurations among servers. By studying these factors, your paper contributes to a deeper understanding of the

decision-making process and emphasizes the importance of considering workload characteristics and server heterogeneity in VM placement strategies.

3. Identification of challenges and open research issues: current paper identifies challenges and open research issues related to VM placement in data centers. These challenges could include optimizing resource utilization, load balancing, minimizing energy consumption, considering fault tolerance, and handling highly dynamic workloads. By highlighting these challenges, current paper encourages further investigation and emphasizes the need for innovative solutions to address these issues in an evolving data center environment.

4. Future directions for researchers to explore: the current paper provides future directions for researchers in the field of VM placement. These directions suggest potential areas that require further investigation, such as integrating SDN technology into the VM placement process, employing machine learning techniques to enhance decision-making, evaluating the impact of other factors (e.g., network latency and security) on placement strategies, and designing advanced algorithms to improve efficiency and scalability. By outlining these future directions, this paper offers guidance for researchers to explore new avenues and contribute to the advancement of VM placement in data centers.

In summary, the novelty of the paper lies in its comprehensive exploration and analysis of thermal-aware virtual machine placement techniques in cloud environments, including emerging paradigms such as statistical machine learning and reinforcement learning-based methods. In addition to providing a thorough review of existing techniques, the paper identifies the factors influencing the placement process, analyzes challenges and open research issues, and provides future directions for researchers.

1.2. **Differentiating static and dynamic approaches in thermal-aware VM placement: A novel perspective**

. It is noteworthy that within extant review papers, such as the one available at [9], the characterization of static methods commonly pertains to instances where the mapping of virtual machines (VMs) remains invariant throughout the entirety of the VM lifecycle, with infrequent recomputation occurring over extended temporal intervals. This static paradigm reflects a strategic deployment approach that prioritizes stability and continuity in VM placements, often guided by historical data or predetermined rules. The prolonged fixed mappings inherent in static methods are conducive to scenarios where the anticipated workload and thermal conditions exhibit relative stability over time. Conversely, dynamic placement methods in these contexts involve the frequent recomputation of VM mappings, contingent upon the prevailing state of the data center. This dynamic paradigm responds promptly to real-time fluctuations in workloads, thermal conditions, and other dynamic factors. The continuous adjustment of VM placements aligns with the dynamic nature of cloud environments, enabling the system to adapt swiftly to varying demands and ensuring optimal resource utilization. This real-time responsiveness distinguishes dynamic methods as a more flexible and adaptive strategy, suitable for cloud environments characterized by unpredictable workloads and rapidly changing operational conditions. As cloud computing continues to evolve, the nuanced understanding of these static and dynamic placement paradigms becomes increasingly crucial for optimizing performance and resource utilization in cloud data centers. However, a closer examination reveals a nuanced understanding that distinguishes our proposed methodology from existing literature. In the subsequent sections of this paper, we delve into a comprehensive exploration of the motivations and challenges associated with thermal-aware virtual machine (VM) placement in cloud data centers. Building upon this foundation, we thoroughly examine the diverse factors that significantly influence the VM placement process, shedding light on critical considerations such as workload characteristics, server heterogeneity, and advanced thermal management techniques. Our analysis extends to an in-depth review of thermal-aware VM placement techniques, where we categorize and discuss static, dynamic, hybrid, machine learning-based, and game theory-based approaches. Subsequently, we delve into the intricacies of evaluating the efficacy of these techniques through dedicated metrics designed for thermal-aware VM placement. Moreover, we provide insights into potential future directions and highlight open research issues, offering a roadmap for researchers to explore in this dynamic and evolving field. The conclusive section synthesizes the key findings, emphasizing the significance of our contributions to the broader discourse on optimizing thermal conditions in cloud data centers.

Following the previous paragraph, the organization of the paper is as follows: section 2 brings about the motivations and challenges of the thermal-aware VM placement strategies. Section 3 discuss about why thermal management is important. Evaluation metrics of the related field are introduced in section 4. As a body of the paper, in section 5 review of the chosen literature is performed. Finally, section 6 conclude the paper.

2. Motivation and challenges of thermal-aware VM placement

Thermal management is a critical aspect of data center operations, as excessive heat can lead to equipment failure and downtime, resulting in significant financial losses. In addition, high temperatures can increase energy consumption and carbon emissions, impacting both the environment and the bottom line. Virtual machine (VM) placement plays a crucial role in managing data center temperature, as the location of VMs on physical servers affects the heat generated by those servers [10]. By strategically placing VMs to balance workload and minimize hot spots, data center operators can optimize resource utilization while maintaining appropriate temperature levels. However, achieving thermal-aware VM placement is not without its challenges. One of the main difficulties is the dynamic nature of cloud computing workloads, which can result in constantly changing thermal profiles. As a result, traditional static placement strategies may not be effective in minimizing hot spots and maintaining temperature levels. Another challenge is the complexity of data center infrastructure, including the cooling system and physical layout. The cooling system must be designed to efficiently remove heat from the data center, and its performance can be impacted by factors such as airflow and humidity [11]. The physical layout of servers can also impact thermal management, as servers located in close proximity can create hot spots and increase overall temperature levels.

Finally, there is a need to balance thermal management with other performance metrics, such as energy consumption and resource utilization. For example, moving VMs to reduce hot spots may result in underutilized resources, leading to increased energy consumption and decreased efficiency. In summary, thermal-aware VM placement is critical for maintaining appropriate temperature levels in data centers and optimizing resource utilization. However, achieving this requires overcoming challenges related to dynamic workloads, complex infrastructure, and balancing performance metrics [12]. Effective solutions must consider all of these factors to achieve optimal thermal management in cloud computing environments. since the VMP problem is NP-hard. there are 4 general approaches to thermal aware VMP. 1- threshold based 2- heuristic 3- metaheuristic and evolutionary 4- machine learning. Threshold-based approaches involve setting temperature thresholds for servers and placing VMs accordingly. This method is simple but may not be effective in dynamic environments. Heuristic approaches use rule-based algorithms to optimize VM placement based on factors such as workload and proximity [13]. These methods are more flexible than threshold-based approaches but may not always find the optimal solution. Metaheuristic and evolutionary approaches use optimization algorithms to find the best VM placement solution. These methods can handle dynamic environments but may be computationally expensive. Machine learning approaches use historical data to predict future workload and temperature patterns and optimize VM placement accordingly. These methods can handle dynamic environments and achieve high accuracy, but require a large amount of data for training. Overall, selecting the best approach for thermal-aware VM placement depends on factors such as data center size, workload characteristics, and available resources. A combination of different approaches may also be necessary to achieve optimal results. An overview of challenges and motivations is given in the following.

2.1. **MOTIVATION**

. 1. Energy efficiency: Data centers consume significant amounts of energy for cooling purposes. By optimizing the placement of VMs based on thermal considerations, energy consumption for cooling can be reduced, leading to improved energy efficiency.

2. Thermal management: High-density server deployments in data centers can result in localized hotspots, which can lead to thermal issues, such as device failures, performance degradation, and reduced equipment lifespan. Thermal-aware VM placement ensures even distribution of heat, improving overall thermal management and mitigating the risks associated with hotspots.

3. Performance optimization: Thermal-aware VM placement can also enhance the performance of VMs by ensuring that they are placed in cooler regions of the data center. This reduces the possibility of throttling and allows for better load balancing, leading to improved overall system performance.

4. Equipment reliability: By considering the thermal profile of servers, VM placement can help prevent thermal stress on hardware components. This, in turn, enhances the reliability and lifespan of the equipment, reducing the probability of failures and associated downtime.

2.2. CHALLENGES

. 1. Accurate thermal modeling: Developing accurate models to characterize the thermal behavior of servers is crucial. This involves understanding heat generation and dissipation patterns within servers and how they are affected by workload variations.

2. Real-time monitoring: Obtaining real-time data on temperature distribution within the data center is a challenge. To make informed placement decisions, it is essential to have up-to-date information on the thermal state of servers.

3. Scalability: As data centers grow in size, the complexity of VM placement increases exponentially. Efficient algorithms and strategies are needed to handle the large-scale nature of data centers and optimize VM placement in a timely manner.

4. Trade-off between conflicting objectives: Thermal considerations need to be balanced with other performance metrics, such as latency, network bandwidth, and power consumption. Finding an optimal trade-off between these conflicting objectives is a non-trivial task in thermal-aware VM placement. According the mentioned challenges, The research seeks to answer several key questions, including: What are the fundamental dimensions and characteristics used to categorize and analyze existing thermal-aware VM placement techniques? How do different methodologies and optimization techniques contribute to thermal-aware VM placement strategies, and what are their respective strengths and limitations? What are the challenges and open research issues in thermal-aware VM placement, and what are the future directions for researchers to explore in this field? How can thermal-aware VM placement strategies enhance the overall performance, reliability, and energy efficiency of fog-cloud infrastructure? What are the selection criteria for comparative analysis of thermal-aware VM placement techniques, and how can these studies contribute to a comprehensive understanding of strategies addressing thermal challenges? These questions form the basis for the comprehensive analysis and

evaluation of thermal-aware VM placement techniques presented in the research paper.

2.3. POTENTIAL RISKS

. the implementation of thermal-aware virtual machine (VM) placement in cloud data centers introduces various potential risks, encompassing aspects related to security, compliance, and more. Here's an elaboration on these potential risks:
1. Security Concerns: Data Vulnerability: Thermal-aware VM placement involves dynamic adjustments in the allocation of virtualized resources. This dynamic nature may introduce security vulnerabilities, such as unauthorized access or data exposure during migration processes. Network Security: VM movements within the data center for thermal optimization purposes may impact network security. The increased traffic generated by these movements could potentially create opportunities for malicious activities or unauthorized access.

2. Compliance Challenges: Data Governance: Adhering to data governance regulations becomes challenging when VMs are dynamically placed to optimize thermal conditions. The movement of VMs may raise concerns related to data sovereignty, especially in regions with strict regulations governing data storage and processing. Industry-Specific Compliance: Various industries have stringent compliance requirements (e.g., healthcare, finance). Thermal-aware VM placement may inadvertently lead to non-compliance with industry-specific regulations regarding data storage, access, and privacy.

3. Operational Risks: Service Disruptions: Frequent VM migrations for thermal optimization purposes may result in service disruptions. If not carefully managed, these disruptions can impact the availability and performance of applications, affecting the overall user experience. Resource Overhead: The continuous monitoring and adjustment of VM placements to manage thermal conditions can introduce additional resource overhead. This may lead to increased energy consumption, counteracting the primary goal of energy efficiency.

4. Interference with Other Optimization Strategies: Conflict with Load Balancing: Thermal-aware VM placement strategies may conflict with load balancing mechanisms. Balancing thermal considerations with other optimization goals, such as load distribution, can be challenging and may lead to suboptimal outcomes. Resource Fragmentation: The dynamic nature of VM placement for thermal management might result in resource fragmentation, impacting resource utilization efficiency and hindering other optimization strategies.

5. Complexity in Implementation: Implementation Challenges: Integrating and implementing thermal-aware VM placement systems can be complex. Configuring systems to consider both thermal dynamics and security measures simultaneously may require sophisticated solutions, leading to potential implementation challenges. Operational Complexity: Managing a thermal-aware

VM placement system introduces operational complexities. IT teams need to adapt to new procedures for monitoring, configuring, and troubleshooting, potentially leading to errors or delays.

6. Environmental Impact: Unintended Environmental Consequences: While the primary goal of thermal-aware VM placement is to improve energy efficiency, there may be unintended consequences. For instance, increased computational loads in specific areas of the data center to optimize thermal conditions may lead to localized hotspots, counteracting the overall goal of efficient thermal management.

7. Scalability Challenges: Scalability Issues: As cloud data centers expand, the scalability of thermal-aware VM placement strategies becomes crucial. Ensuring that these strategies can effectively scale to meet the demands of larger and more complex data center infrastructures without compromising performance is a significant challenge. To mitigate these risks, thorough risk assessments, adherence to compliance standards, robust security measures, and careful consideration of operational implications are essential. Additionally, ongoing research and advancements in thermal-aware VM placement technologies aim to address and alleviate these risks.

3. The importance of thermal management in data centers

Thermal management is a critical aspect of data center operations, as excessive heat can lead to equipment failure and downtime, resulting in significant financial losses. When data centers get too hot, it can cause damage to hardware components such as processors, memory, and storage devices. This can result in equipment failure and downtime that can be costly for businesses. In addition, high temperatures can increase energy consumption and carbon emissions, impacting both the environment and the bottom line. Therefore, it is crucial to maintain appropriate temperature levels in data centers to ensure optimal performance, reliability, and energy efficiency. When performing virtual machine (VM) placement in data centers, several typical thermal constraints need to be considered. These constraints include temperature limits, airflow and ventilation management, heat dissipation, and redundancy/failover mechanisms. Data centers aim to maintain equipment within recommended temperature ranges, ensure proper airflow to prevent hotspots, balance heat dissipation to avoid overburdening cooling infrastructure, and have backup systems in case of cooling failures. However, these thermal constraints can vary across data center environments. Factors such as location, infrastructure, and design play a role in these variations. For instance, data centers in hotter regions may face more significant cooling challenges than those in cooler climates. The infrastructure and design of a data center, including the cooling technology used, can influence thermal constraints. Different equipment types and utilization levels contribute to variations as well. Considering these variations, data centers must tailor their VM placement strategies to their specific thermal constraints.

This involves monitoring and modeling temperatures, optimizing airflow management, simulating the impact of VM placement on cooling efficiency, and adapting strategies accordingly. By addressing these thermal constraints effectively, data centers can ensure proper cooling, prevent equipment failure, and optimize energy usage.

Virtual machine (VM) placement plays a crucial role in managing data center temperature, as the location of VMs on physical servers affects the heat generated by those servers [14]. By strategically placing VMs to balance workload and minimize hot spots, data center operators can optimize resource utilization while maintaining appropriate temperature levels. This means that VMs should be placed in a way that evenly distributes workloads across servers and avoids overloading specific servers. This approach helps to prevent hot spots from forming and maintains a consistent temperature across the data center [15]. By optimizing VM placement for thermal management, businesses can reduce energy consumption, improve equipment reliability, and minimize downtime. Achieving thermal-aware VM placement is not without its challenges. One of the main difficulties is the dynamic nature of cloud computing workloads, which can result in constantly changing thermal profiles [16]. As a result, traditional static placement strategies may not be effective in minimizing hot spots and maintaining temperature levels. This means that data center operators must continuously monitor and adjust VM placement to maintain optimal thermal conditions. Another challenge is the complexity of data center infrastructure, including the cooling system and physical layout. The cooling system must be designed to efficiently remove heat from the data center, and its performance can be impacted by factors such as airflow and humidity. The physical layout of servers can also impact thermal management, as servers located in close proximity can create hot spots and increase overall temperature levels. Finally, there is a need to balance thermal management with other performance metrics, such as energy consumption and resource utilization. For example, moving VMs to reduce hot spots may result in underutilized resources, leading to increased energy consumption and decreased efficiency. Therefore, data center operators must consider all performance metrics when making decisions about VM placement. By optimizing VM placement for both thermal management and performance, businesses can achieve the best possible outcomes in terms of efficiency, reliability, and cost-effectiveness.

since the VMP problem is NP-hard [17]. there are 4 general approaches to thermal aware VMP. 1- threshold based 2- heuristic 3- metaheuristic and evolutionary 4- machine learning. A detailed explanation for these four general approaches is as follows:

1. Threshold-based approach: This approach is based on setting predefined temperature thresholds for servers and VMs. Once a threshold is exceeded, the placement of VMs is adjusted to reduce the temperature. This approach is relatively simple and easy to implement, but it does not account for the dynamic nature of workloads and may lead to suboptimal placements.

2. Heuristic approach: This approach uses a set of rules or guidelines to determine VM placement based on factors such as workload balance and proximity to other VMs. Heuristics are often based on intuition or past experience and can be effective in reducing hot spots and maintaining temperature levels.

3. Metaheuristic and evolutionary approach: These approaches use optimization algorithms to find optimal VM placements based on a set of criteria, including thermal management, resource utilization, and performance metrics [18]. Metaheuristic algorithms, such as simulated annealing or tabu search, use a set of rules to iteratively improve VM placements. Evolutionary algorithms, such as genetic algorithms or particle swarm optimization, use genetic operators to evolve solutions over time. These approaches can be highly effective in finding optimal solutions but can be computationally expensive.

4. Machine learning approach: This approach uses machine learning algorithms to learn from historical data and predict optimal VM placements. Machine learning models can be trained on data such as workload patterns, server temperatures, and energy consumption to predict future placements that balance thermal management with other performance metrics. This approach can be highly effective but requires large amounts of data and computational resources for training [19].

Overall, each approach has its strengths and weaknesses, and the choice of approach will depend on the specific needs and constraints of the data center. A combination of approaches may also be used to achieve the best possible outcomes in terms of thermal management and performance. It identifies several challenges in implementing energy-efficient resource management in virtualized HPC cloud datacenters. One challenge is the conflicting objectives of maximizing energy efficiency and utilization of cloud datacenter resources while avoiding undesired thermal hotspots. Another challenge is the need to maintain service level agreements (SLAs) for application runtime while consolidating virtual machines onto fewer servers. Traditional server consolidation techniques can violate SLAs due to greater resource contention at higher utilization levels. A third challenge is the dynamic nature of HPC workloads, which can lead to unpredictable changes in resource demand and thermal conditions. Finally, there is a need for proactive and thermal-aware resource management techniques that consider multiple pairwise tradeoffs and use intelligent virtual machine migrations to avoid potential thermal hotspots. Proactive thermal-aware resource management in virtualized HPC cloud datacenters involves using thermodynamic models and real-time measurements to capture the complex thermodynamic phenomena of heat generation and heat extraction [20]. The goal is to predict the future temperature map of the datacenter to enable proactive thermal-aware datacenter management decisions. The designed technique involves a proactive and thermal-aware virtual machine consolidation (involving allocations as well as migrations) technique that maximizes computing resource utilization, minimizes datacenter energy consumption for computing, and improves the efficiency of heat extraction. The technique considers potential thermal hotspots and avoids them through intelligent virtual machine migrations. The results show that this technique achieves better results in terms of energy efficiency, resource utilization, and thermal management compared to other state-of-the-art techniques. Specifically, this technique achieves up to 30% energy savings compared to traditional server consolidation techniques while maintaining SLAs for application runtime. This technique reduces the number of idle or lightly-loaded servers by up to 50% which maximizes computing resource utilization.

Using clouds for HPC applications have several benefits. One benefit is the abstraction of nearly-unlimited computing resources through the elastic use of federated resource pools (virtualized datacenters). This can provide HPC users with on-demand access to computing resources without having to invest in expensive hardware or maintain their own datacenters. Another benefit is the ability to scale up or down computing resources based on workload demand, which can improve resource utilization and reduce energy consumption. The approach proposed in [21] enhances the understanding of resource allocation and workload distribution in a federated cloud system. By incorporating thermal-aware VM placement strategies as discussed in this survey paper, the efficiency of big data management can be improved, allowing for better utilization of resources and reduced energy consumption. Clouds can also provide a more flexible and dynamic environment for HPC applications, allowing users to experiment with different configurations and software stacks without having to worry about hardware constraints. Finally, clouds can provide a more cost-effective solution for HPC applications by reducing capital expenditures (CAPEX) and operating expenses (OPEX) associated with traditional supercomputing clusters or grids. Overall, the potential benefits of using clouds for HPC applications include improved resource utilization, reduced energy consumption, greater flexibility and scalability, and lower costs.

4. Evaluation metrics for thermal-aware VM placement

Evaluating the effectiveness of thermal-aware virtual machine (VM) placement strategies is imperative for assessing their impact on data center performance and energy efficiency. A range of metrics has been devised to quantify and analyze the outcomes of different placement methodologies. These metrics provide valuable insights into the thermal dynamics, energy consumption, and overall efficiency of the virtualized infrastructure. In this section, we explore and categorize the key evaluation metrics commonly employed in the context of thermal-aware VM placement, offering a nuanced understanding of the performance criteria considered in the literature.

4.1. Temperature Distribution Metrics:

4.1.1. *Inlet Temperature:*
Definition: The temperature of the air entering the servers [22]. Significance: Provides insights into the thermal conditions at the source, influencing the overall temperature distribution within the data center.

4.1.2. *Heat Recirculation Effect:*
Definition: Measures the extent to which heated air is recirculated within the data center. Significance: Indicates the efficiency of the VM placement strategy in minimizing heat recirculation, contributing to a more uniform temperature distribution [39].

4.2. **Energy Consumption Metrics:**

4.2.1. *Total Energy Consumption:*
Definition: The overall energy consumed by the data center, encompassing both computational tasks and thermal management. Significance: Offers a holistic view of energy usage, aiding in the assessment of the environmental impact and operational costs [23].

4.2.2. *Power Usage Efficiency (PUE):*
. Definition: The ratio of total energy consumption to the energy consumed by IT equipment, providing an assessment of data center energy efficiency. Significance: Reflects the overall energy utilization efficiency, considering both computational tasks and associated cooling requirements.

4.3. **Resource Utilization Metrics:**

4.3.1. *Server Utilization:*
Definition: Measures the extent to which computing resources are utilized on individual servers. Significance: A high server utilization indicates efficient resource allocation, while low utilization may signify underutilized resources or potential overloading.

4.3.2. *VM Consolidation Ratio:* Definition: The ratio of active VMs to the total VM capacity, indicating the degree of consolidation. Significance: Evaluates how effectively VMs are consolidated onto a minimal number of servers, contributing to resource optimization and reduced energy consumption.

4.4. **Performance Metrics:**

4.4.1. *SLA Violation Metric:*

Definition: Measures the extent to which Service Level Agreements (SLAs) are violated, indicating the system's ability to meet performance guarantees. Significance: Offers insights into the impact of VM placement on system responsiveness and user experience, prioritizing the adherence to predefined performance standards [24].

4.4.2. *Throughput:*

Definition: The rate at which tasks or transactions are processed by the system. Significance: Measures the efficiency of VM placement in supporting high-performance computing demands. Definition: The rate at which tasks or transactions are processed by the system. Significance: Measures the efficiency of VM placement in supporting high-performance computing demands.

4.5. **Scalability Metrics:**

4.5.1. *Scalability Factor:*

Definition: Measures the ability of the VM placement strategy to scale with increasing workloads or data center size. Significance: Assesses the adaptability of the strategy to evolving demands, ensuring sustained efficiency in dynamic cloud environments.

In the subsequent sections, we will employ these comprehensive metrics to evaluate and compare existing thermal-aware VM placement techniques. By scrutinizing these metrics, we aim to provide a thorough assessment of the strengths and limitations of various strategies, contributing to the ongoing discourse on optimizing thermal conditions in cloud data centers.

5. THERMAL-AWARE VM PLACEMENT TECHNIQUES – LITERATURE REVIEW

In this section, as the key section of the paper, the previous related literature is investigated. To organize the analysis, a well explained classifications is needed to get a lucrative insight from various proposed methods that is presented in the field of thermal-aware VM placement. This section involves categorizing and analyzing existing literature based on key dimensions and characteristics. The taxonomy employed in this review paper is anchored in six fundamental dimensions: temporal, spatial, methodology and optimization techniques, workload characteristics, energy efficiency metrics, and real-world implementation. Each dimension serves as a lens through which we analyze and evaluate the diverse strategies devised to address the complex challenges associated with thermal-aware VM placement.

5.1. CATEGORIZATION BASIS

. The categorization basis is pivotal for several reasons. Firstly, it provides a structured framework that allows for a holistic examination of the multifaceted approaches deployed in thermal-aware VM placement. By systematically organizing the literature along distinct dimensions, we gain deeper insights into the variations, strengths, and limitations of different strategies. This structured approach not only facilitates a comprehensive understanding of the field but also aids in identifying gaps and areas for future exploration. Moreover, the necessity of categorization lies in its ability to distill the vast and diverse body of research into discernible patterns and trends. As the thermal-aware VM placement landscape continues to expand, a well-defined taxonomy becomes instrumental in navigating the wealth of information. Researchers, practitioners, and stakeholders alike can benefit from a clear delineation of strategies based on temporal considerations, spatial scopes, and various other critical factors. In this section of this review paper, we will delve into each dimension of the taxonomy, unraveling the intricacies of thermal-aware VM placement strategies. By systematically examining the temporal evolution, spatial considerations, methodological intricacies, workload dynamics, energy efficiency metrics, and real-world implementations, we aim to provide a comprehensive and insightful survey that contributes to the collective understanding of thermal-aware virtual machine placement in cloud data centers. Figure 1 illustrates the categorization basis of this review.

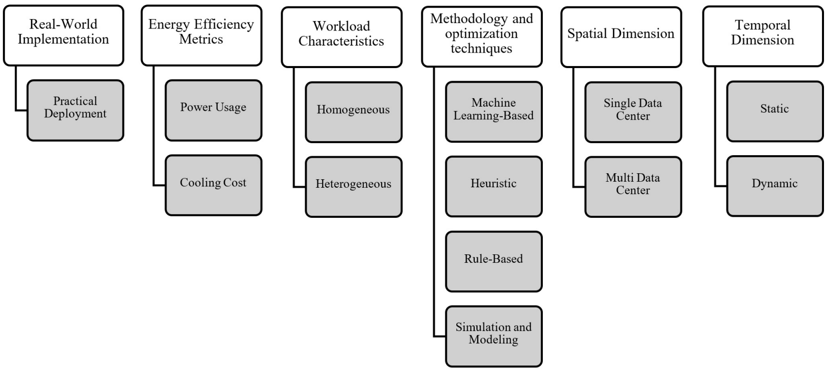

Figure 1. Categorization basis of this review.

Temporal dimension: The temporal dimension in thermal-aware virtual machine (VM) placement is delineated into static and dynamic approaches. Static methods involve fixed placements or infrequent recomputation based on historical data, with an innovative twist introducing adaptability within the ostensibly fixed mapping. Dynamic methods, on the other hand, dynamically adjust VM placements in response to real-time data center conditions. Our taxonomy adds a proactive dynamic element, emphasizing a forward-looking approach to anticipate and prevent thermal issues. This nuanced understanding of the temporal dimension enhances the adaptability and responsiveness of VM placement strategies to the dynamic nature of thermal dynamics, contributing a novel perspective to the field. Spatial Dimension: This dimension classifies thermal-aware virtual machine (VM) placement studies based on the scope of deployment. Single Data Center focuses on VM placement within a singular data center, while Multi-Data Center extends the analysis to distributed or multi-cloud environments. This categorization enables a nuanced exploration of challenges and optimizations within different spatial contexts. Methodology and Optimization Techniques: Here, studies are categorized based on the methodologies employed for thermal-aware VM placement. Machine Learning-Based Approaches leverage algorithms for predictive optimization, Heuristic and Rule-Based Methods rely on deterministic guidelines, and Simulation and Modeling use theoretical or simulated models for analysis. This classification provides insights into the diverse approaches used for optimizing VM placements. Workload Characteristics: Examining the characteristics of workloads, this dimension distinguishes between Homogeneous Workloads, where VM placements cater to uniform workloads, and Heterogeneous Workloads, where diverse workloads require specialized placement strategies. Understanding how VM placements adapt to workload diversity is crucial for practical implementation. Energy Efficiency Metrics: This dimension assesses the impact of VM placements on energy efficiency. Power Consumption focuses on minimizing power usage, while Cooling Costs evaluate the influence on overall energy efficiency by considering the cooling requirements. These metrics provide a comprehensive view of the energy implications associated with different VM placement strategies. Real-world Implementations: This dimension explores the practical aspects of thermal-aware VM placement. Practical Deployments discuss real-world applications and case studies, while Challenges and Limitations critically analyze obstacles encountered during implementation. This provides valuable insights into the feasibility and challenges associated with deploying thermal-aware VM placement in operational environments.

5.2. **SELECTION CRITERIA FOR COMPARATIVE ANALYSIS**

. Choosing studies for comparative analysis in the realm of thermal-aware virtual machine (VM) placement requires a judicious and systematic approach. To ensure a comprehensive and insightful review, we employ the following selection criteria: Relevance to Thermal-Aware VM Placement: Studies must explicitly focus on thermal considerations in VM placement within cloud data centers. This ensures that the selected works directly contribute to the understanding of strategies addressing thermal challenges. Publication Quality: Preference is given to studies published in reputable peer-reviewed journals, conferences, or academic platforms. This criterion ensures the inclusion of high-quality research that has undergone rigorous scrutiny within the academic community.

Variety in Methodologies: To offer a well-rounded analysis, we include studies employing diverse methodologies, encompassing static, dynamic, hybrid, machine learning-based, and game theory-based approaches. This approach allows for a thorough examination of the spectrum of strategies used for thermal-aware VM placement.

5.3. STUDY ANALYSIS

. The proposed method in [25] is dynamic, as it involves proactive thermal-aware virtual machine consolidation and migration based on real-time resource utilization and temperature data. The system continuously monitors the datacenter environment and makes decisions on virtual machine allocation and migration to optimize resource utilization and minimize energy consumption while ensuring that the recommended operating temperature is not exceeded. the technique can be applied to multiple datacenters, allowing for even greater resource utilization and energy savings across a larger infrastructure. The proposed method is based on simulation and modeling, as it involves the design and validation of a heat-imbalance model to predict future temperature trends for optimal resource allocation in datacenters. The proposed method primarily focuses on addressing the challenges posed by heterogeneous workloads in datacenters. the authors implemented and validated their method in real-world settings. They conducted feasibility studies and proof-of-concept measurement-based experiments at the NSF Center for Cloud and Autonomic Computing (CAC) at Rutgers University.

Figure 2 shows an envisioned cross-layer approach for managing HPC datacenters [25].

Proactive thermal-aware virtual machine consolidation technique involves both allocations and migrations. It minimizes datacenter energy consumption for computing by turning off unused servers after workload consolidation, which saves energy. It improves the efficiency of heat extraction by consolidating virtual machines onto hotter server aisles, which allows for more efficient heat extraction by the CRAC system. One drawback is that traditional server consolidation can violate service level agreements (SLAs) in terms of application runtime due to greater resource contention at higher utilization levels. Another drawback is that traditional server consolidation can lead to thermal hotspots, which can cause overheating and reduce the reliability and lifespan of computing equipment. This technique overcomes these drawbacks by being proactive and thermal-aware, considering multiple pairwise tradeoffs, and using intelligent virtual machine migrations to avoid potential thermal hotspots. In contrast to the proactive methodology employed in [25], the methodology proposed in the work [22] utilizes a heuristic-based optimization technique founded on a total power consumption model. The authors of the aforementioned study employ a fuzzy controller in their proposed methodology to enhance the performance of the optimizer. By accounting for the heat recirculation effect and mitigating the over-consolidation side effect, the authors achieve a reduction

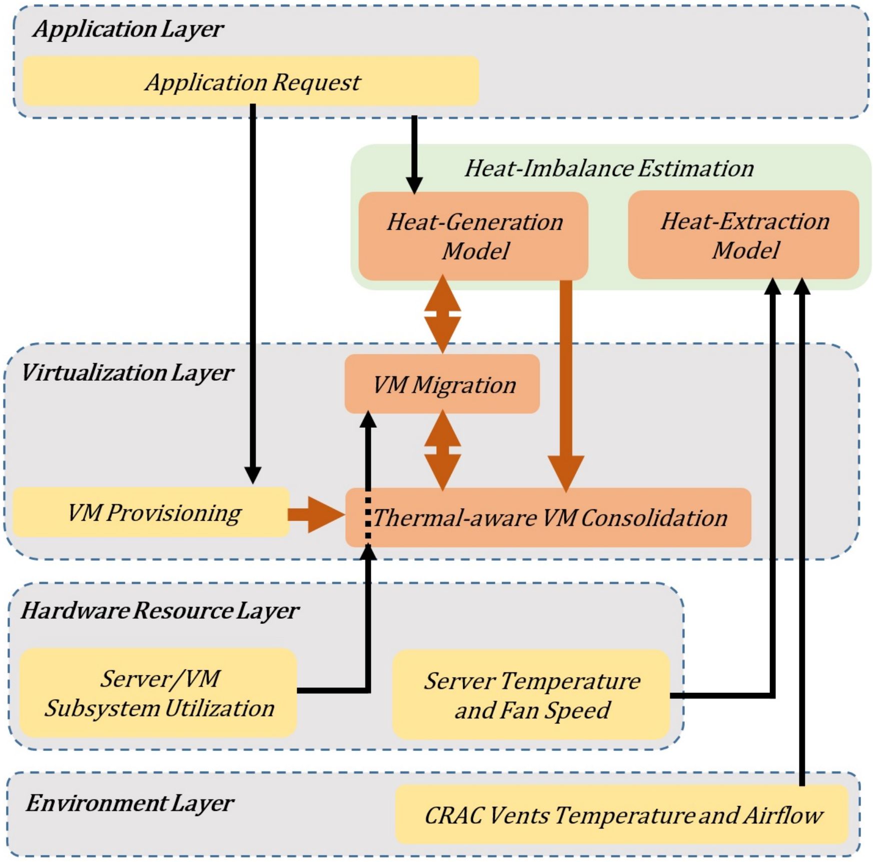

Figure 2. Envisioned cross-layer approach for managing HPC data centers proposed in [25].

in energy consumption and cooling costs while mitigating excessive VM migration. Consequently, the proposed method endeavors to uphold throughput levels while concurrently minimizing energy consumption. Although the primary focus of the methodology is on a singular data center, its scalability to encompass multiple data centers is acknowledged. However, it remains a challenge for the optimizer to effectively navigate the expanded solution space associated with multiple data centers. The authors validate their proposed approach through simulation, leveraging a simulator that accommodates two distinct workloads to demonstrate the efficacy of their methodology.

A fundamental drawback of model-based approaches lies in the inherent limitations of the model itself, presenting a substantial challenge concerning the delicate balance between complexity and computability. Conversely, model-free methodologies, which engage in online interaction with the environment, circumvent the need for predefined models to derive optimized policies. Instead, they leverage reinforcement learning algorithms to glean insights from environmental feedback. Despite the promise inherent in this paradigm, model-free methods face two significant challenges. Firstly, they contend with random, delayed feedback from the environment, diminishing the experiential efficiency of the agent. Secondly, the intricate nature of the environment necessitates modeling in a state-action space manner, thereby constraining the scalability of reinforcement learning solutions. In a recent study published in [26], the authors propose a novel approach that employs spatial, as opposed to traditional temporal, coding to address these challenges. Their methodology is grounded in a hierarchical tree structure of physical states, which concurrently serves as an agent processing bed, adopting a decentralized approach. The authors incorporate energy consumption and cooling efficiency metrics into the reward function of the learning algorithm. While the proposed method lacks support from real-world implementation, extensive simulation experiments under heterogeneous workloads were conducted. The scalability of this approach, crucial for supporting federated cloud data centers, is emphasized. Unlike decentralized decision-making architectures, Software Defined Networking (SDN) Based approaches provides a centralized view of the network, allowing administrators to monitor and control network resources more efficiently. In the context of power-aware VM placement, SDN enables fine-grained control over network traffic and resource allocation, enabling the identification of underutilized hosts and redistributing VMs to consolidate workloads.

5.4. **VM allocators for cloud data centers**

. VM allocators are designed to be a building block of a Software Defined Networking (SDN) orchestrator. The designed allocators use Fuzzy Logic, Single and Multi-Objective optimization algorithms to allocate VM resources following two different policies, namely Best Fit and Worst Fit, corresponding to consolidation and spreading strategies respectively. A set of Virtual Machine (VM) allocators for Cloud Data Centers (DCs) is designed that perform the joint allocation of computing and network resources. The designed allocators use Fuzzy Logic, Single and Multi-Objective optimization algorithms to allocate VM resources following two different policies, namely Best Fit and Worst Fit, corresponding to consolidation and spreading strategies respectively [27]. For each server, the allocators choose the network path that minimizes electrical power consumption, evaluated according to a precise model specifically designed for network switches. The study uses power consumption as the energy metric. Specifically, it considers the power consumption of network devices, including access switches, aggregation switches, and core switches. They did not provide any real world scenario. Simulation tests have been carried out to evaluate the performance of the allocators in terms of number of allocated VMs for each policy. They did not focus on the workload characteristics. Figure 3 shows resource allocator algorithm.

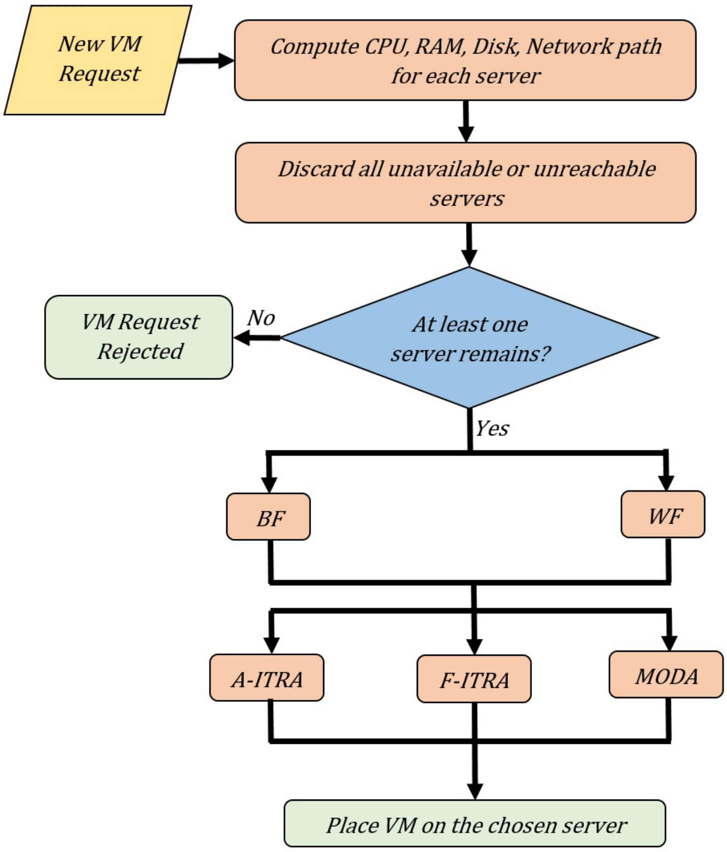
New VM Request
Compute CPU, RAM, Disk, Network path for each server
Discard all unavailable or unreachable servers
At least one server remains?
No
VM Request Rejected
Yes
BF
WF
A-ITRA
F-ITRA
MODA
Place VM on the chosen server

Figure 3. VM placement algorithm presented in [27].

While there are typically unused servers with different performance levels in such data centers, conventional thermal management techniques do not take advantage of these resources to cool down hot servers, the proposed method in [28] is a location-aware adaptive DTM (Dynamic Thermal Management) technique for heterogeneous data centers. The technique adaptively exploits external as well as internal computing resources available in a rack, considering application characteristics and server locations. The proposed technique employs three methods: (1) L-MigrationPM, which migrates a VM to another physical machine in a rack with different performance considering its location, (2) MigrationCore, which migrates VMs among CPU cores in the physical machine, and (3) a DVFS-based method. The technique samples on-chip temperature of cores in each physical machine periodically and estimates the performance impact of VM migrations, including performance degradation due to the physical machine migration and/or core migration of VMs. Depending on the estimated performance impact of VM migrations, the technique adaptively employs the above three methods. The proposed method is dynamic, as it adaptively exploits external and internal computing resources based on real-time measurements and estimates of the performance impact of VM migrations. The technique dynamically selects among the three methods (L-MigrationPM, MigrationCore, and DVFS-based method) based on the on-chip temperature of cores in each physical machine and the estimated performance impact of VM migrations. the proposed DTM technique for heterogeneous data centers considers cooling costs. It employs a modeling approach to estimate the performance impact of VM migrations and select the most appropriate method for thermal management. The proposed DTM technique is designed to support heterogeneous workloads in data centers. The proposed approach for heterogeneous data centers is primarily focused on managing thermal aspects and optimizing performance within a single data center environment. Figure 4 shows the thermal cross interference among distributed nodes.

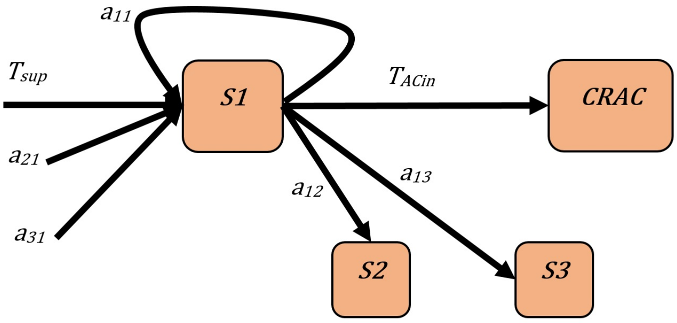

Figure 4. Thermal cross interference among different nodes in [28].

Previous methods used to optimize energy consumption in cloud data centers have focused on specific dimensions of energy usage, such as server usage efficiency, network usage efficiency, and cooling efficiency. The proposed method in [29] differs from previous studies in several key aspects: Comprehensive Approach: Unlike some existing works that focus on optimizing only one or two dimensions of energy consumption, the proposed method addresses cooling, server, and network energy consumption to provide a more holistic optimization approach. Consideration of Non-IT Resources: The proposed method takes into account the energy consumption of non-IT resources such as the cooling system and heat-recirculation effect, which is not commonly addressed in previous VMP strategies. Two-Step Algorithm: The use of the two-step SAG algorithm, combining Simulated Annealing and Greedy algorithms, sets this method apart from other VMP strategies and contributes to its effectiveness in reducing total energy consumption. An overview of the method is demonstrated in figure 5.

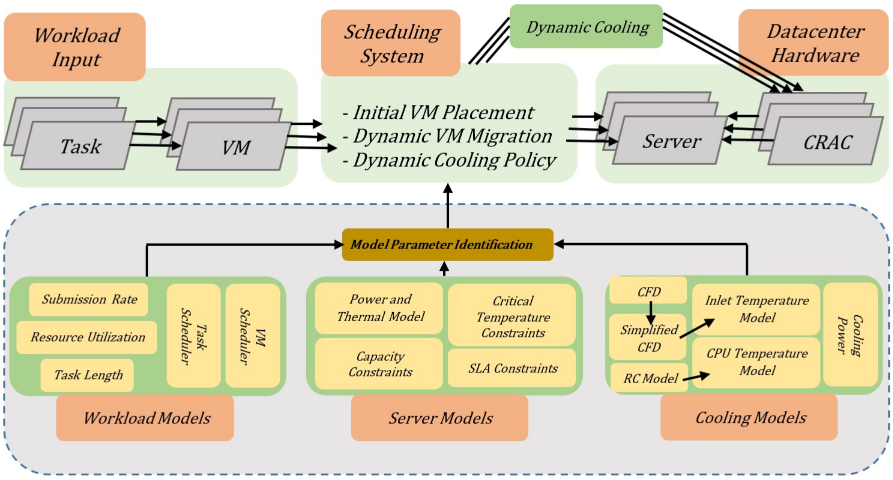

FIGURE 5. VM placement and dynamic cooling approach presented in [29].

The explained method considers multiple energy metrics, including server energy consumption, network energy consumption, and cooling energy consumption. They use a comprehensive model that includes heat recirculation effect to minimize energy consumption. Understanding and addressing heat recirculation is crucial for optimizing the energy efficiency of data centers. Authors in [30] proposed VMP (Virtual Machine Placement) strategy takes into account heat recirculation in data centers through the development of a heat-recirculation-aware VM placement strategy, known as SABA. This strategy is designed to minimize energy consumption and avoid hot spots in data centers by considering the impact of heat recirculation on the overall energy efficiency. SABA incorporates two key features to address heat recirculation: 1. Achieving Thermal Balance: The VMP strategy aims to achieve an approximate optimal thermal balance within the data center, thereby reducing the energy consumption of the cooling system and improving cooling efficiency. 2. Energy-Saving through Server Activation: SABA reduces the number of active servers required for VM tasks, thereby yielding significant energy savings and contributing to the overall energy efficiency of the data center. Furthermore, the strategy utilizes a simulated annealing based algorithm to lower the energy consumption of physical resources, taking into account the heat recirculation coupled with multiple physical resource allocations. By integrating heat-recirculation-aware considerations into the VM placement strategy, SABA effectively addresses the impact of heat recirculation on energy consumption and cooling efficiency in data centers, making it a distinctively energy-efficient VMP strategy. The effectiveness of the strategy and algorithm was validated through a series of simulation-based experiments using real-world data center workloads. The research and experiments primarily focused on individual data centers, but the principles and strategies developed, including the consideration of heat recirculation and energy-efficient VM placement, have the potential to be extended to multiple data centers. As data center operations become larger and more complex, the demand for energy-efficient strategies grows, making the proposed approach adaptable to address the challenges faced by multiple data centers. In [31], the authors also use simulated annealing as a heuristic optimizer. The method proposed in the study is called Burst-aware and Thermal-efficient Virtual Machine Placement (BTVMP). It is designed to address the challenges of bursty workloads and energy efficiency in cloud data centers. The BTVMP method consists of two main components: 1. Split-and-Recombine Algorithm: BTVMP leverages the Split-and-Recombine (SAR) algorithm to classify incoming workloads into two categories: ordinary workloads and key workloads. These workloads are further classified based on their execution time into three types: long-execution-time, medium-execution-time, and short-execution-time. SAR prioritizes key workloads while monitoring the execution rate of ordinary workloads to prevent starvation. The method then recombines the key and ordinary workloads into virtual machines (VMs) for future scheduling, ensuring that a fixed proportion of computing resources is allocated to execute ordinary workloads to avoid the starvation problem. Additionally, short-term workloads are given precedence over other ordinary workloads to attain VM resources for execution. 2. Enhanced Simulated Annealing Algorithm (ESA): BTVMP utilizes an enhanced simulated annealing algorithm to assign VMs on physical servers in an energy-efficient manner. The ESA algorithm initializes an initial solution with a relatively large value near the top of the rack to speed up convergence. It iteratively explores the search space by changing the direction of its new solution towards other directions of the current solution. ESA also judiciously assigns VMs on a set of active nodes while powering off low-utilization nodes to enhance the utilization of computing nodes. Energy metrics include total energy consumption, computing energy consumption, cooling energy consumption, and thermal efficiency. In [35] authors go to a finer grain of placement. They try to transform VMP into a deep level of multicore servers. They believe to optimize VM placement in cloud data centers, incorporating awareness of core scheduling into the VM placement algorithm must be considered. This involves mapping virtual CPUs (vCPUs) onto physical CPUs (pCPUs), resulting in an optimization problem that combines VM placement and core scheduling. The authors used constraint programming techniques to solve the VM placement problem in cloud data centers. They argued that a simplified model of scheduling issues within a single physical machine (PM) should be taken into account during VM placement. They showed how constraint programming can be used to effectively solve this problem and improve the placement of VMs. This method provides a possible problem formulation for incorporating core scheduling into VM placement, which involves mapping vCPUs onto pCPUs. Incorporating awareness of core scheduling into virtual machine (VM) placement algorithms can lead to several advantages in cloud data centers. It can increase energy efficiency: By reducing the number of PMs needed to support a given workload, the designed method can also help reduce energy consumption in cloud data centers. Incorporating core scheduling awareness into VM placement algorithms can lead to significant improvements in performance, cost, and energy efficiency in cloud data centers. The evaluation compares the designed algorithm to several baseline algorithms that do not incorporate core scheduling awareness. The results show that the designed algorithm outperforms the baseline algorithms in terms of both performance and energy efficiency. Specifically, the designed algorithm achieves up to 30% improvement in application performance and up to 20% reduction in energy consumption compared to the baseline algorithms. There are some limitations of the evaluation, such as the use of synthetic workloads and assumptions about PM and VM characteristics, but overall suggests that the results support the effectiveness of incorporating core scheduling awareness into VM placement algorithms in cloud data centers. Another limitation lies behind the scalability. Using fine grain scheduling always skeptical for performance maintenance during scaling up operation. By minimizing both IT equipment power consumption and cooling system power consumption, the energy efficiency of data centers can be optimized. In [36], authors use a thermo-electrical DC model that defines the relation between server utilization, server room ambient temperature, and cooling system load needed to dissipate the corresponding heat. Additionally, a set of approximation algorithms and heuristics are designed to solve the resulting NP-hard thermal-aware consolidation problem. Finally, simulation results on some predefined scenarios show these techniques improve DC energy consumption with 5% up to 20% compared to the well-known First-Fit algorithm by deploying VMs on servers with low thermal influence. A DC optimization methodology is designed that aims at reducing the overall DC energy consumption by consolidating the VMs in a thermal-aware manner. The methodology is based on the DC components thermo-electrical models that relate the server utilization due to VM deployment to the ambient server room temperature and to the cooling system energy consumption. A set of approximation algorithms and heuristics is designed to solve the resulting NP-hard thermal-aware consolidation problem.

Authors in [37] propose a greedy-based algorithm that considers the energy consumption of both servers and cooling systems, as well as the workload distribution across different servers. Additionally, it provides a comprehensive evaluation of the designed solution through simulations and experiments on real-world Cloud datacenters. A method called GRANITE is designed, which is a greedy-based algorithm for virtual machine scheduling in Cloud datacenters. GRANITE takes into account both the energy consumption of servers and cooling systems, as well as the workload distribution across different servers. The algorithm works by iteratively selecting a server with the highest potential for energy savings and migrating virtual machines to that server. This process continues until no further energy savings can be achieved. Figure 6 shows holistic cloud data center models for total-energy-aware VM scheduling

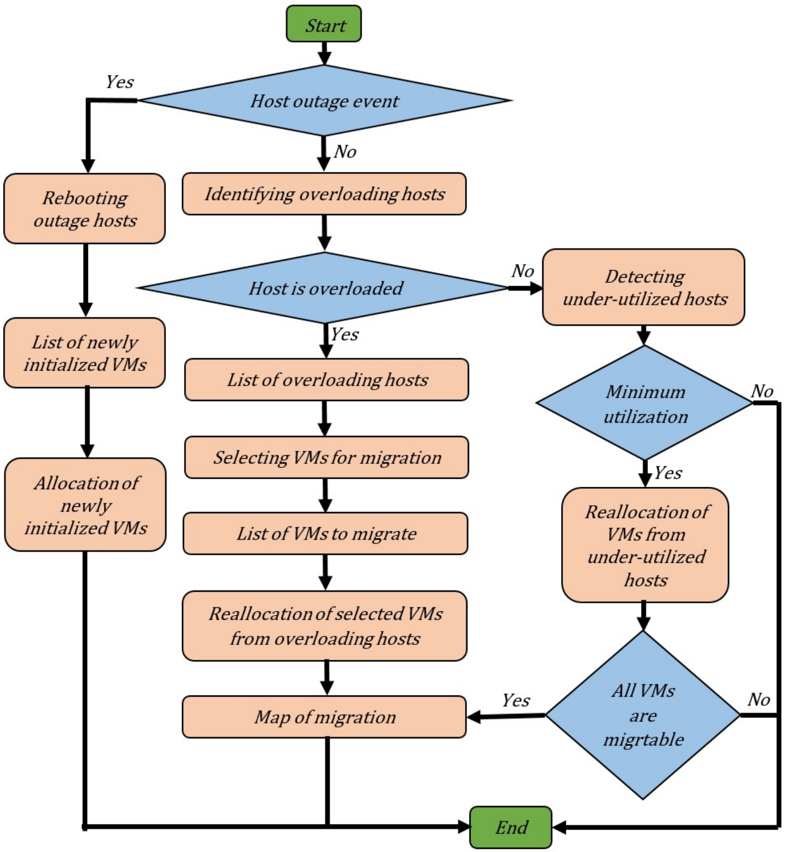
Start
Host outage event
Yes
No
Rebooting outage hosts
Identifying overloading hosts
Host is overloaded
No
Detecting under-utilized hosts
Yes
List of newly initialized VMs
List of overloading hosts
Minimum utilization
No
Selecting VMs for migration
Yes
Allocation of newly initialized VMs
List of VMs to migrate
Reallocation of VMs from under-utilized hosts
Reallocation of selected VMs from overloading hosts
Yes
All VMs are migrtable
No
Map of migration
End

FIGURE 6. Holistic cloud datacenter models for total-energy-aware VM scheduling proposed in [37].

The impact of host temperature on server reliability in cloud data centers is significant. A study has shown that the variability in temperature of a host has a negative impact on server reliability [32]. The larger the variation in host temperature, the greater the likelihood of an outage occurring. High operating temperature is often believed to be the major cause of hardware failures, but the variability in temperature has a stronger negative impact on server reliability. The method proposed in [38] considers the variability in host temperature as a migration criterion to avoid outage incidents and achieve better VM consolidations. It aims to minimize energy consumption and maximize resource utilization while reducing the number of server outage incidents due to fluctuations in host temperature. The mechanism takes into account both host power consumption and temperature to achieve a balance between energy saving and SLA (Service Level Agreement) violations. It also incorporates a Markov model to predict future CPU usages of physical hosts and VMs, reducing the number of migrations needed in the long run. Extensive simulation experiments conducted on CloudSim demonstrate the promising performance of the proposed mechanism in energy saving and outage avoidance. Among the selected power-based methods, destination hosts were chosen based on the power consumption. In contrast, in the designed mechanism, the host with the lowest CoV (Coefficient of Variation) after migration was chosen. Experiment results that highlight its advantages in energy conservation, overload avoidance, and outage avoidance. Figure 7 shows the virtual machine consolidation process of the mentioned study. Simulation results show that the designed mechanism outperforms existing power-based methods in terms of energy conservation and resource utilization while maintaining acceptable levels of temperature and performance. The evaluation results suggest that the designed mechanism is a comprehensive solution for efficient energy and temperature management in cloud data centers.

A multi-objective VM placement algorithm must be able to jointly consider energy efficiency and other criteria such as SLA and thermal effects. Many approaches use migration strategy to compensate imprecise decision-making policies. In [33], researchers focus on optimizing energy consumption in cloud data centers by considering the thermal effects of intensive workloads on IT equipment. The multi-objective optimization model used in this research is called MOPFGA (Multi-objective algorithm based on Pathfinder Algorithm and Genetic Algorithm). It combines the classic MOPFA and GA enhanced by OBL (Opposition Based Learning) for fast convergence and avoidance of local optimum. MOPFGA considers three main objectives: migration cost, energy consumption, and heat recirculation around server racks. The algorithm uses a workload model to represent the computational resource constraints in the system and minimize the migration cost target. It also proposes a heat-recirculation model to ensure that the computer room can run within thermal constraints. Finally, a power consumption model considering computational and cooling energy consumption is proposed according to the above model. The proposed thermal-aware VMP strategy is designed to address the challenges of virtual machine placement in single data centers. They use simulations using real-world traces to evaluate their methods.

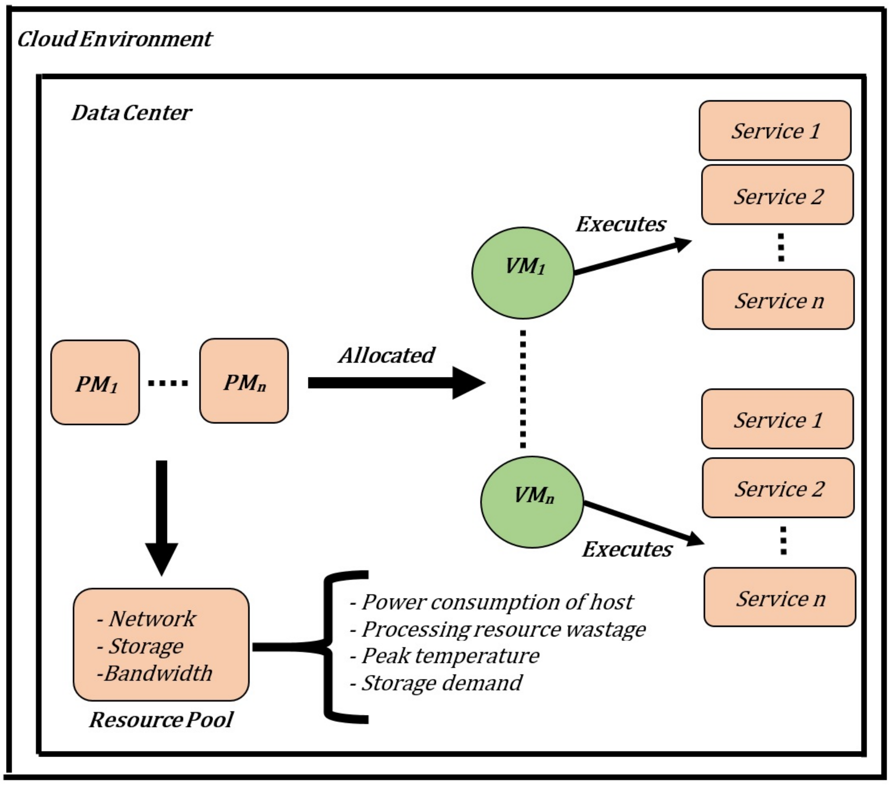

Figure 7. Virtual machine consolidation process presented in [38].

The proposed resource scheduling method presented in [34], aims to reduce energy consumption by optimizing total energy consumption and proactively preventing hot spots from a global perspective. The method includes four phases where the virtual machine scheduler uses an improved ant colony algorithm to find appropriate target hosts for virtual machines based on server temperature and utilization status obtained in real-time. The scheduling system minimizes energy consumption by dynamically consolidating virtual machines, and the decision of the scheduling system is based on the models established in the paper, including the computing system power consumption model, cooling system power consumption model, and temperature model. The primary optimization goal is to minimize the total power consumption of the data center while avoiding hot spots and simultaneously reducing energy consumption without compromising the quality of service. The approach utilizes energy metrics such as the computing system power consumption model, cooling system power consumption model, and total power consumption to optimize energy usage within the data center. By dynamically consolidating virtual machines and making decisions based on real-time server temperature and utilization status, the method aims to minimize total power consumption, prevent hot spots, and reduce energy consumption without compromising the quality of service. The evaluation of the proposed method is not based on real-world implementation. The main focus is on the scheduling within a data center and they did not mention any support of multiple data centers. The proposed resource scheduling method based on thermal management does not appear to be limited to any specific workload.

An online scheduling algorithm based on GRASP meta-heuristic is utilized by authors in [8] for dynamic VM consolidation to reduce energy consumption in computing systems while avoiding the creation of local hotspots that can negatively impact energy consumption and system reliability. The efficiency of the designed algorithm is validated through extensive experiments using real workload traces in a simulated environment, and its superiority is demonstrated by comparing it to several baseline algorithms. A new approach is provided to dynamic virtual machine consolidation that addresses the challenges of energy consumption and system reliability in cloud data centers. The online scheduling algorithm based on GRASP meta-heuristic for dynamic virtual machine consolidation to optimize the distribution of workloads (VMs) in cloud data centers. The algorithm is designed to holistically and proactively prevent hotspots by efficiently distributing workloads between computing and cooling subsystems. The energy metrics considered in the study cover both computational and cooling energy. The simulation scenarios fed by PlanetLab workloads that is considered to be a computationally intensive one. It can be inferred that the proposed algorithm was not evaluated against heterogeneous workloads. The primary scheduling intention of the method is within a single data center.

Researchers in [40] provide a thermal-aware Virtual Machine (VM) allocation heuristic based on the genetic algorithm, and discusses the importance of considering both computing and cooling energy consumption in heterogeneous data centers. To achieve a higher energy saving in VMs, it's important to create a formal definition of optimal thermal-aware VM allocation by considering both computing and cooling energy consumption and providing a novel heuristic based on a genetic algorithm to obtain a near-optimal solution in less computing time while designing a trade-off between the power-aware consolidation techniques and thermal-aware load balancing approaches. The paper considers both computing and cooling energy consumption as the key energy metrics for minimizing the total energy consumption of a heterogeneous data center. The computing energy consumption is calculated based on the CPU utilization of each chassis, while the cooling energy consumption is calculated based on the Coefficient of Performance (CoP) of the Computer Room Air Conditioning (CRAC) unit. The total energy consumption of a data center is defined as the sum of chassis energy and the cooling energy of the CRAC unit. Their evaluations are based on simulated experiments in modified version of CloudSim platform. The formulation should be able to mathematically formulate the problem of how to allocate Virtual Machines in a thermal-aware manner among the computing nodes in order to minimize the total energy consumption of the data center (the formulated problem is named MITEC) [40]. Heuristic approach to solve the MITEC problem based on a genetic algorithm. Furthermore, the performance of the designed heuristic is compared with thermal-aware greedy algorithms and power-aware VM allocation heuristics. A thermal-aware Virtual Machine (VM) allocation heuristic is designed based on a genetic algorithm to minimize the total energy consumption of a heterogeneous data center. The problem of minimizing the total energy consumption of a heterogeneous data center (MITEC) is formulated as a non-linear integer optimization problem. Figure 8 shows the interaction of various entities in cloud data center for resource management.

This method provides an effective solution for managing energy consumption in data centers while considering both computing and cooling energy consumption. Overall, the evaluation results demonstrate the effectiveness of the designed method in reducing energy consumption and costs in data centers while considering both computing and cooling energy consumption.

To evaluate such approaches one way is to integrate thermal behavior models into existing platforms like CloudSim. In [41] authors proposed a framework named ThermoSim based on CloudSim which enables researchers to evaluate their methods reliably [43]. ThermoSim framework is used to simulate and model thermal-aware resource management for cloud computing environments. ThermoSim is a novel framework that uses a lightweight RNN-based deep learning model to predict temperature characteristics of cloud hosts, which can be used to optimize energy consumption and temperature simultaneously in resource-constrained cloud environments. ThermoSim extends the CloudSim toolkit and shows thermal-aware and utilization-based approaches for scheduling resources, which can enhance the performance of cloud data centers. The designed scheduling approaches equipped with efficient energy and thermal-aware policies can improve QoS parameters such as energy consumption, SLA violation rate, number of VM migrations, and temperature. ThermoSim has low overhead for resource management due to its lightweight RNN-based deep learning predictor for temperature characteristics of cloud hosts. ThermoSim framework is designed using three well-known energy-aware and thermal-aware resource scheduling techniques, which demonstrates its effectiveness in optimizing energy consumption and temperature simultaneously in resource-constrained cloud environments. The evaluation results showed that ThermoSim outperformed the existing thermal-aware simulator in terms of energy consumption and temperature. The designed scheduling approaches equipped with efficient energy and thermal-aware policies also improved QoS

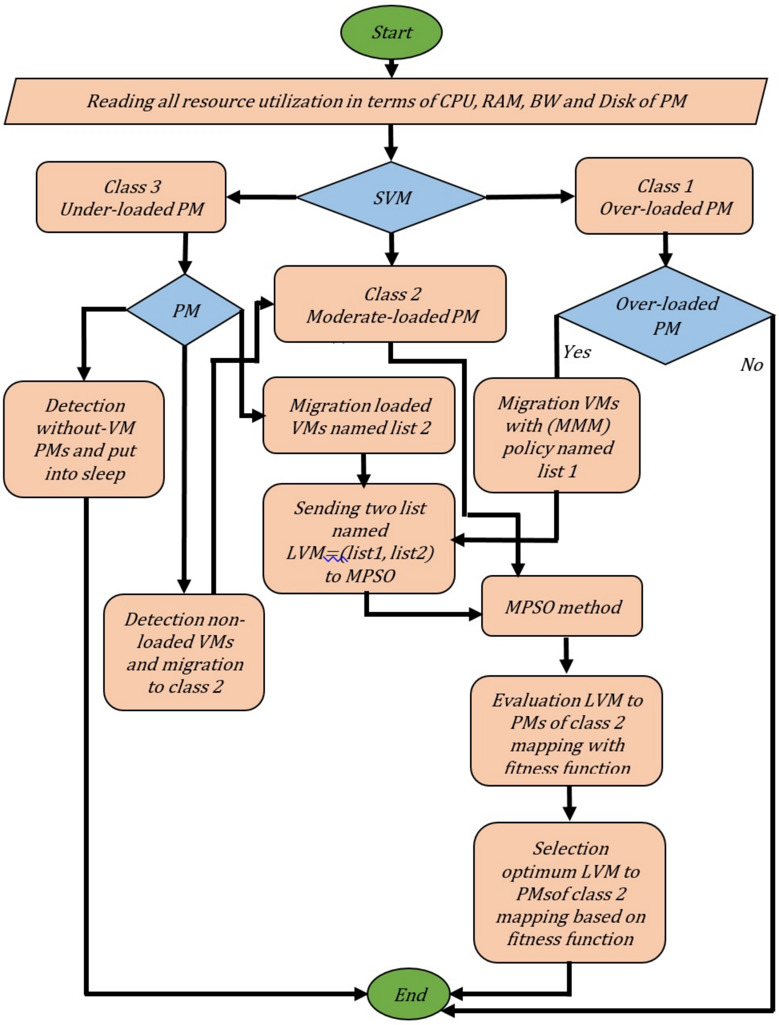

Figure 8. Interaction of various entities in cloud data center for resource management proposed in [40].

parameters such as SLA violation rate and number of VM migrations. ThermoSim is an effective framework for optimizing energy consumption and temperature simultaneously in resource-constrained cloud environments.

Recently many researchers inspiring machine learning techniques are trying to use historical data to improve the inefficiencies of existing model. In [42] authors develop a platform called EVMC based on statistical machine learning to cluster the physical machines. Then using a metaheuristic algorithm perform allocation. EVMC is a meta-heuristic and machine learning method called EVMC, which is an effective method in contrast to other algorithms for the VMC in the CCSs. It considers all resources of PMs in VMC for meeting the energy efficiency in the CCSs and presents the metrics and a technique for considering a trade-off between power consumption and other QoS in the CCSs. This technique presents a formal verification architecture for the energy-aware EVMC method by using the PAT tool. It considers proposing a series of formal experiments for verification of the divergence-freeness, deadlock-freeness, and reachability. Therefore, an innovative approach is used to address energy wastage and low server utilization challenges in cloud data centers by presenting an energy-aware virtual machines consolidation (EVMC) method that optimizes energy consumption while guaranteeing quality of service (QoS) through machine learning techniques, modified minimization of migration approach, and modified particle swarm optimization [42]. Additionally, it presents formal verification architecture and experiments to ensure its effectiveness. Energy-aware virtual machines consolidation (EVMC) method is designed to optimize energy consumption while ensuring quality of service (QoS) in cloud computing systems. The EVMC method comprises three main components: 1. Support vector machine classification method based on the utilization rate of all resources of physical machines (PMs) that is used for PM detection in terms of the amount' load. 2. Modified minimization of migration approach which is used for virtual machine (VM) selection. 3. Modified particle swarm optimization which is implemented for VM placement. The functional requirements of the EVMC method are evaluated using formal methods and non-functional requirements through simulation. A formal verification architecture is designed for the energy-aware EVMC method using the PAT tool and shows a series of formal experiments to verify its effectiveness in terms of divergence-freeness, deadlock-freeness, and reachability. EVMC method provides better energy efficiency and utilization of resources while ensuring quality of service (QoS) in cloud computing systems compared to other existing methods. Figure 9 shows the flowchart of the EVMC method [42]. Most VM placement approaches ignore the load of tasks that may arrive to virtual machines to execute. On the other hand, task scheduler methods often presume that virtual machine already placed. But in [43], researchers proposed a holistic approach called JTSVMP which considers both aspects of scheduling in cloud data centers. JTSVMP stands for "joint task scheduling and virtual machine placement" in cloud data centers. This method is designed to address the challenges of VM allocation in cloud computing environments, such as workload balancing, energy efficiency, and resource utilization. In this method, joint task scheduling and virtual machine placement (JTSVMP) as modeled one co-optimization problem. Metaheuristic optimization algorithms (MOA) are used to solve this problem and plan a schedule defining not only the task to VM mapping but also the VM to physical host (PH) mapping. The performance of the co-optimization process is evaluated by comparing it with two different scenarios:

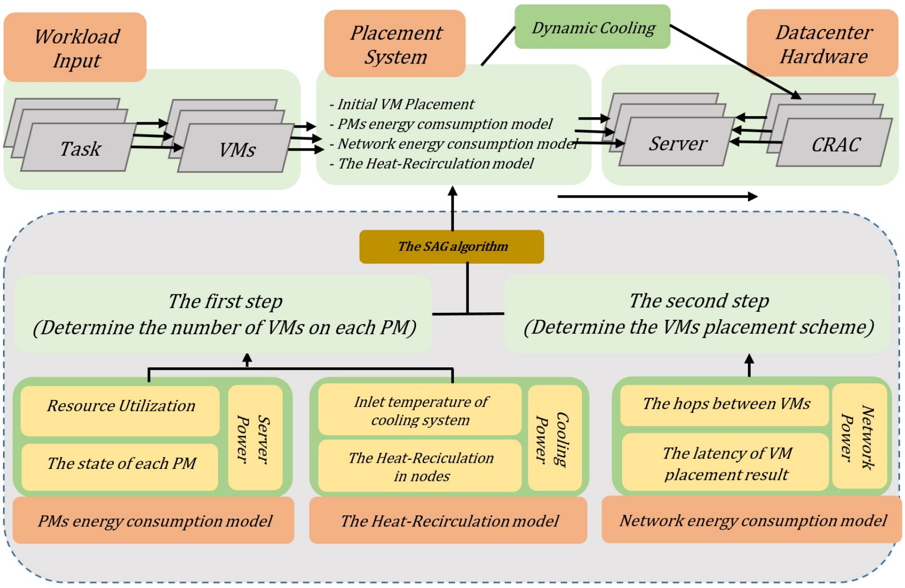
Workload Input
Task
VMs
Placement System
- Initial VM Placement
- PMs energy comsumption model
- Network energy consumption model
- The Heat-Recirculation model
Dynamic Cooling
Datacenter Hardware
Server
CRAC
The SAG algorithm
The first step
(Determine the number of VMs on each PM)
The second step
(Determine the VMs placement scheme)
Resource Utilization
The state of each PM
Server Power
Inlet temperature of cooling system
The Heat-Reciculation in nodes
Cooling Power
The hops between VMs
The latency of VM placement result
Network Power
PMs energy consumption model
The Heat-Recirculation model
Network energy consumption model

Figure 9. Structure of the EVMC method presented in [42]
.

task scheduling algorithms and integration co-optimization of task scheduling and VM placement using MOAs. Three different MOAs, namely basic glowworm swarm optimization (GSO), moth-flame glowworm swarm optimization (MFGSO), and genetic algorithm (GA) are used to solve the JTSVMP problem. Extensive experimental evaluations are conducted using a cloud data center simulator to demonstrate the effectiveness of this approach compared to other state-of-the-art methods. The results showed that this method outperformed existing approaches in terms of solution quality and computational efficiency [43]. JTSVMP stands for "joint task scheduling and virtual machine placement" in cloud data centers. This method is designed to address the challenges of VM allocation in cloud computing environments, such as workload balancing, energy efficiency, and resource utilization. In this method, joint task scheduling and virtual machine placement (JTSVMP) as modeled one co-optimization problem. Metaheuristic optimization algorithms (MOA) are used to solve this problem and plan a schedule defining not only the task to VM mapping but also the VM to physical host (PH) mapping. The performance of the co-optimization process is evaluated by comparing it with two different scenarios: task scheduling algorithms and integration co-optimization of task scheduling and VM placement using MOAs. Three different MOAs, namely basic glowworm swarm optimization (GSO), moth-flame glowworm swarm optimization (MFGSO), and genetic algorithm (GA) are used to solve the JTSVMP problem. Extensive experimental evaluations are conducted using a cloud data center simulator to demonstrate the effectiveness of this approach compared to other state-of-the-art methods. The results showed that this method outperformed existing approaches in terms of solution quality and computational efficiency. This method offers a promising solution to the challenges of VM allocation in cloud computing environments and has potential applications in various types of cloud data centers and workloads. Thermal-aware VMP approaches can be very lucrative in many emerging technologies. For example, In [43] Authors highlight the importance of cooperative concurrency control mechanisms for IoT transaction processing in a fog-cloud computing environment. By integrating thermal-aware VM placement strategies into the proposed approach, as examined in this survey paper, the authors can enhance the system's thermal management, optimize resource allocation, and improve the overall performance and reliability of the fog-cloud infrastructure.

In summation of this section, Table 1 and table 2 provide a comprehensive listing of the scrutinized studies, delineating their respective features and characteristics.

6. Conclusion and future directions

The increasing demand for cloud computing services has led to the development of large-scale data centers with a high number of servers and virtual machines. These data centers consume a significant amount of energy and require efficient thermal management to ensure optimal performance and reliability. Thermal-aware virtual machine placement has emerged as a promising approach to address these challenges by optimizing the placement of virtual machines based on their thermal characteristics. Thermal-aware virtual machine placement techniques can be broadly categorized into five approaches: static, dynamic, hybrid, machine learning-based, and game theory-based. Static approaches rely on pre-defined rules and heuristics to place virtual machines in a data center. Dynamic approaches, on the other hand, use real-time data to make placement decisions. Hybrid approaches combine both static and dynamic techniques to achieve better results. Machine learning-based approaches use machine learning algorithms to learn from historical data and make predictions about future placement decisions. Game theory-based approaches model the placement problem as a game between virtual machines and data center resources. Each approach has its advantages and limitations. For example, static approaches are simple and easy to implement but may not be optimal in dynamic environments. Dynamic approaches can adapt to changing conditions but may incur high overheads. Hybrid approaches aim to balance the

Table 1. Summary of literature review.

Ref	**Year**	**Temporal dimension**	**Spatial dimension**	**Optimization technique**	**Energy efficiency metrics**	**Real-world implementation simulation**	**Workload characteristics**
25	2015	Dynamic	Multiple data centers	Modeling	Cooloing cost-computational energy-thermal violation	real-world implementation	heterogeneous
22	2022	Dynamic	Single data centers	Metaheuristic	Computational energy- cooling energy-heat recirculation	Simulation	heterogeneous
26	2023	Dynamic	Multiple data centers	Machine Learning	Computational Power- thermal Violation	Simulation	heterogeneous
27	2016	Static	Single data centers	Heuristic, Fuzzy logic	Computational power, network bandwidth usage	Simulation	heterogeneous
28	2021	Dynamic	Single data centers	Heuristic	Computational power, cooling cost	Simulation	heterogeneous

strengths of both static and dynamic techniques but may require more complex algorithms. Machine learning-based approaches can learn from historical data and make accurate predictions but may require a large amount of training data. Game theory-based approaches can model complex interactions between virtual machines and data center resources but may be computationally expensive. This review paper provides a comprehensive overview of the state-of-the-art techniques for thermal-aware virtual machine placement and highlights the advantages and limitations of each approach. It aims to assist researchers and practitioners in selecting the appropriate approach for their specific requirements. Future research directions have also been identified, including the development of more efficient algorithms, the integration of multiple approaches, and the evaluation of thermal-aware virtual machine placement in real-world data centers. Researchers and practitioners aiming to advance the field of thermal-aware virtual machine placement should adopt an interdisciplinary approach, collaborating with experts from diverse fields like computer science, thermal engineering, and data center management. Staying informed about emerging technologies, particularly in machine learning and data analytics, can enhance the efficiency and adaptability of proposed strategies. Prioritize real-world validation of solutions, engaging with industry partners and

Table 2. Continue from table 1.

Ref	Year	Temporal dimension	Spatial dimension	Optimization technique	Energy efficiency metrics	Real-world implementation simulation	Workload characteristics
29	2021	Static	Multiple data centers	Heuristic	Computational power, cooling cost, network usage, heat recirculation effect	Simulation	heterogeneous
30	2021	Dynamic	Single data centers	Metaheuristic	Computational power, cooling cost, heat recirculation effect	Simulation	heterogeneous
31	2023	Dynamic	Single data centers	Metaheuristic	Computational energy, cooling energy, thermal efficiency	real-world implementation	heterogeneous
35	2016	Static	Single data centers	Rule-based	Computational energy	Simulation	heterogeneous
36	2016	Static	Single data centers	Heuristic	Computational energy, thermal efficiency	Simulation	heterogeneous
37	2017	Dynamic	Multiple data centers	Heuristic	Computational energy, cooling energy	real-world implementation	heterogeneous
38	2018	Dynamic	Multiple data centers	Rule-based	Computational energy, thermal stability	Simulation	heterogeneous
33	2023	Dynamic	Single data centers	Metaheuristic	Computational power, cooling cost, heat recirculation effect	Simulation	heterogeneous
34	2023	Dynamic	Single data centers	Metaheuristic	Computational energy, cooling energy	Simulation	heterogeneous
8	2018	Dynamic	Single data centers	Metaheuristic	Computational power, cooling energy, heat recirculation effect	Simulation	heterogeneous
42	2020	Static	Single data centers	Metaheuristic	Computational power, cooling energy	Simulation	heterogeneous
43	2020	Dynamic	Multiple data centers	Machine Learning	Computational power, cooling energy, heat recirculation effect	Simulation	heterogeneous
44	2021	Dynamic	Single data centers	Machine Learning, Metaheuristic	Computational power, cooling energy, heat recirculation effect	real-world implementation	heterogeneous
45	2021	Dynamic	Single data centers	Metaheuristic	Computational power, workload balancing	Simulation	heterogeneous

deploying strategies in operational data centers. Addressing energy efficiency metrics, security concerns, and ethical considerations is crucial for developing responsible and impactful thermal-aware VM placement methodologies. Embracing open-source practices, contributing to standardization efforts, and exploring multi-data center environments will further contribute to the robustness and scalability of research findings. Continuous learning through participation in conferences and workshops will keep researchers abreast of emerging trends and foster collaboration within the research community.

References

[1] Versick, D., Tavangarian, D. (2013). The CÆSARA architecture for power and thermal-aware placement of virtual machines. Paper presented at the 2013 International Green Computing Conference Proceedings.

[2] Sun, H., Stolf, P., Pierson, J.-M., Da Costa, G. J. S. C. I., Systems. (2014). Energy-efficient and thermal-aware resource management for heterogeneous datacenters. 4(4), 292-306.

[3] Chaudhry, M. T., Ling, T. C., Manzoor, A., Hussain, S. A., Kim, J. J. A. C. S. (2015). Thermal-aware scheduling in green data centers. 47(3), 1-48.

[4] Mhedheb, Y., Streit, A. (2016). Energy-efficient task scheduling in data centers. Paper presented at the International Conference on Cloud Computing and Services Science.

[5] Liu, X., Gu, H., Zhang, H., Liu, F., Chen, Y., Yu, X. J. M., Microsystems. (2017). Energy-Aware on-chip virtual machine placement for cloud-supported cyber-physical systems. 52, 427-437.

[6] Ananthi, M. S. S. D. B. Virtual Machine Management for Cloud Data Center to Avoid Security Issues.

[7] Salimian, L., Safi-Esfahani, F. J. I. J. o. G., Computing, U. (2018). Energy-efficient placement of virtual machines in cloud data centres based on fuzzy decision making. 9(4), 367-384.

[8] Kaur, A., Singh, V., Gill, S. S. (2018). The future of cloud computing: opportunities, challenges and research trends. Paper presented at the 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on.

[9] Masdari, M., Nabavi, S. S., Ahmadi, V. J. J. o. N., Applications, C. (2016). An overview of virtual machine placement schemes in cloud computing. 66, 106-127.

[10] Van Damme, T., De Persis, C., Tesi, P. J. I. T. o. C. S. T. (2018). Optimized thermal-aware job scheduling and control of data centers. 27(2), 760-771.

[11] Reddy, M. A., Ravindranath, K. J. A. A. I. (2020). Virtual machine placement using JAYA optimization algorithm. 34(1), 31-46.

[12] Ahmed, K., Yoshii, K., Tasnim, S. (2019). Thermal-aware power capping allocation model for high performance computing systems. Paper presented at the 2019 International Conference on Computational Science and Computational Intelligence (CSCI).

[13] Nath, K. R., Sreeram, G., Lavanya, D., Kiran, U., Rajesh, P. J. I. J. o. A. S., & Technology. (2019). Efficient virtual machine placement in data center. 28(16), 580-587.

[14] Qiu, Y., Jiang, C., Wang, Y., Ou, D., Li, Y., & Wan, J. J. E. (2019). Energy aware virtual machine scheduling in data centers. 12(4), 646.

[15] Omer, S., Azizi, S., Shojafar, M., Tafazolli, R. J. J. o. s. a. (2021). A priority, power and traffic-aware virtual machine placement of IoT applications in cloud data centers. 115, 101996.

[16] Tang, Q., Mukherjee, T., Gupta, S. K., Cayton, P. (2006). Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters. Paper presented at

the 2006 Fourth international conference on intelligent sensing and information processing.

[17] Fernandez de La Vega, W., Lueker, G. S. J. C. (1981). Bin packing can be solved within 1+ e in linear time. 1(4), 349-355.

[18] Blum, C., Roli, A. J. A. c. s. (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. 35(3), 268-308.

[19] Yi, D., Zhou, X., Wen, Y., Tan, R. J. I. T. o. P., Systems, D. (2020). Efficient compute-intensive job allocation in data centers via deep reinforcement learning. 31(6), 1474-1485.

[20] Liao, D., Sun, G., Yang, G., Chang, V. J. F. G. C. S. (2018). Energy-efficient virtual content distribution network provisioning in cloud-based data centers. 83, 347-357.

[21] Stergiou, C. L., Psannis, K. E., Gupta, B. B. (2021). InFeMo: flexible big data management through a federated cloud system. ACM Transactions on Internet Technology (TOIT), 22(2), 1-22.

[22] Aghasi, A., Jamshidi, K., Bohlooli, A. (2022). A thermal-aware energy-efficient virtual machine placement algorithm based on fuzzy controlled binary gravitational search algorithm (FC-BGSA). Cluster Computing, 1-19.

[23] Kumar, D., Kulshrestha, S. (2018). Energy Efficient Task Scheduling in Cloud Data Center. International Journal of Distributed Cloud Computing, 6(2).

[24] Chen, R., Liu, B., Lin, W., Lin, J., Cheng, H., Li, K. (2023). Power and thermal-aware virtual machine scheduling optimization in cloud data center. Future Generation Computer Systems, 145, 578-589.

[25] Lee, E. K., Viswanathan, H., Pompili, D. J. I. T. o. C. C. (2015). Proactive thermal-aware resource management in virtualized HPC cloud datacenters. 5(2), 234-248.

[26] Aghasi, A., Jamshidi, K., Bohlooli, A., Javadi, B. (2023). A decentralized adaptation of model-free Q-learning for thermal-aware energy-efficient virtual machine placement in cloud data centers. Computer Networks, 224, 109624.

[27] Portaluri, G., Adami, D., Gabbrielli, A., Giordano, S., Pagano, M. (2016). Power consumption-aware virtual machine allocation in cloud data center. Paper presented at the 2016 IEEE Globecom Workshops (GC Wkshps).

[28] Kim, Y. G., Kim, S. Y., Choi, S. H., Chung, S. W. (2021). Thermal-aware adaptive VM allocation considering server locations in heterogeneous data centers. Journal of Systems Architecture, 117, 102071.

[29] Feng, H., Deng, Y., Li, J. (2021). A global-energy-aware virtual machine placement strategy for cloud data centers. Journal of Systems Architecture, 116, 102048.

[30] Feng, H., Deng, Y., Zhou, Y., Min, G. (2021). Towards heat-recirculation-aware virtual machine placement in data centers. IEEE Transactions on Network and Service Management, 19(1), 256-270.

[31] Li, J., Deng, Y., Wang, R., Zhou, Y., Feng, H., Min, G., Qin, X. (2023). BTVMP: A Burst-Aware and Thermal-Efficient Virtual Machine Placement Approach for Cloud Data Centers. IEEE Transactions on Services Computing.

[32] El-Sayed, N., Stefanovici, I. A., Amvrosiadis, G., Hwang, A. A., Schroeder, B. (2012, June). Temperature management in data centers: Why some (might) like it hot. In Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems (pp. 163-174).

[33] Liu, B., Chen, R., Lin, W., Wu, W., Lin, J., Li, K. (2023). Thermal-aware virtual machine placement based on multi-objective optimization. The Journal of Supercomputing, 1-28.

[34] Mao, L., Chen, R., Cheng, H., Lin, W., Liu, B., Wang, J. Z. (2023). A resource scheduling method for cloud data centers based on thermal management. Journal of Cloud Computing, 12(1), 1-18.

[35] Mann, Z. Á. J. I. T. o. C. (2016). Multicore-aware virtual machine placement in cloud data centers. 65(11), 3357-3369.

[36] Marcel, A., Cristian, P., Eugen, P., Claudia, P., Cioara, T., Anghel, I., Ioan, S. (2016). Thermal aware workload consolidation in cloud data centers. Paper presented at the 2016 IEEE 12th international conference on intelligent computer communication and processing (ICCP).

[37] Li, X., Garraghan, P., Jiang, X., Wu, Z., Xu, J. J. I. T. o. p., systems, d. (2017). Holistic virtual machine scheduling in cloud datacenters towards minimizing total energy. 29(6), 1317-1331.

[38] Wang, J. V., Cheng, C. T., Tse, C. K. J. S. P., Experience. (2019). A thermal-aware VM consolidation mechanism with outage avoidance. 49(5), 906-920.

[39] Ilager, S., Ramamohanarao, K., Buyya, R. J. C., Practice, C., Experience. (2019). ETAS: Energy and thermal-aware dynamic virtual machine consolidation in cloud data center with proactive hotspot mitigation. 31(17), e5221.

[40] Akbari, A., Khonsari, A., Ghoreyshi, S. M. J. E. (2020). Thermal-aware virtual machine allocation for heterogeneous cloud data centers. 13(11), 2880.

[41] Gill, S. S., Tuli, S., Toosi, A. N., Cuadrado, F., Garraghan, P., Bahsoon, R., . . . Software. (2020). ThermoSim: Deep learning based framework for modeling and simulation of thermal-aware resource management for cloud computing environments. 166, 110596.

[42] Zolfaghari, R., Sahafi, A., Rahmani, A. M., Rezaei, R. J. S. P., Experience. (2022). An energy-aware virtual machines consolidation method for cloud computing: Simulation and verification. 52(1), 194-235.

[43] Al-Qerem, A., Alauthman, M., Almomani, A., Gupta, B. B. (2020). IoT transaction processing through cooperative concurrency control on fog–cloud computing environment. Soft Computing, 24, 5695-5711.

Sajed Dadashi
Orcid number: 0000-0002-2192-315X
Department of Computer Engineering
Islamic Azad University
Roudsar, Iran
Email address: sajed.dadashi@gmail.com

Ali Aghasi
Orcid number: 0000-00002-5954-0090
Department of Computer Engineering
University of Isfahan
Isfahan, Iran
Email address: ali.aghasi62@gmail.com