Resource Allocation Model for Energy-Efficient Virtual Machine Placement in Data Centers

KABIR SHOLAGBERU AHMED[1], OLUSHOLA DAMILARE ODEJOBI[2]
[1, 2]Independent Researcher Lagos Nigeria

Abstract- The rapid expansion of cloud computing has intensified the energy demands of modern data centers, raising concerns over sustainability, operational costs, and environmental impact. Virtualization has emerged as a key enabler for improving resource utilization, yet the placement of virtual machines (VMs) across physical hosts remains a complex challenge, particularly when balancing performance guarantees with energy efficiency. Inefficient VM allocation leads to server overutilization, resource fragmentation, and increased power consumption, ultimately undermining both cost efficiency and carbon reduction efforts. This proposes a resource allocation model designed to optimize energy-efficient VM placement in large-scale data centers. The model integrates principles of virtualization, dynamic workload management, and energy-aware computing to achieve multi-objective optimization across three dimensions: resource utilization, energy efficiency, and quality of service (QoS) compliance. Core mechanisms include heuristic and AI-driven VM placement algorithms, predictive workload modeling, and dynamic migration strategies that minimize energy waste while preserving service-level agreements (SLAs). The model also incorporates thermal-aware scheduling to reduce cooling demands and policy-driven orchestration to align allocation decisions with sustainability and compliance standards. Evaluation metrics for the framework span both technical and environmental domains, including CPU, memory, and network utilization, SLA violation rates, VM migration costs, total energy consumption, and power usage effectiveness (PUE). By consolidating workloads intelligently and leveraging predictive allocation, the model reduces the number of active servers, curtails cooling requirements, and supports carbon footprint reduction without sacrificing application performance. The proposed resource allocation model offers strategic implications for cloud service providers and enterprises, enabling them to achieve cost optimization, operational resilience, and alignment with green computing objectives. Ultimately, energy-efficient VM placement represents a critical pathway toward sustainable, scalable, and environmentally responsible cloud infrastructure.

Keywords: Resource Allocation, Energy Efficiency, Virtual Machine Placement, Data Centers, Cloud Computing, Workload Consolidation, Power Consumption Optimization, Dynamic Resource Management, Server Utilization, Green Computing, Thermal Management, Energy-Aware Scheduling, Performance Optimization

I. INTRODUCTION

The unprecedented growth of cloud computing has transformed the global digital ecosystem, enabling enterprises, governments, and individuals to access computing resources on demand (Battleson *et al.*, 2016; Yang *et al.*, 2017). This paradigm shift has been powered largely by large-scale data centers, which serve as the backbone of modern digital infrastructure. These facilities host thousands of servers, networking devices, and storage systems, supporting workloads that range from simple web hosting to artificial intelligence (AI) and high-performance computing (HPC) (Sikeridis *et al.*, 2017; Stoica *et al.*, 2017). However, this rapid expansion has come with a critical cost: rising energy consumption. Data centers are now among the largest industrial consumers of electricity worldwide, accounting for a significant portion of global power usage and contributing directly to carbon emissions (Dayarathna *et al.*, 2015; Shehabi *et al.*, 2016). The energy footprint of these infrastructures is driven not only by the servers themselves but also by associated cooling systems and network operations, making efficiency a pressing challenge in the era of sustainable computing (Gough *et al*., 2015; Lent, 2016).

A key contributor to inefficiency in data centers lies in virtual machine (VM) placement (Alshaer, 2015; Tomer, 2017). Yet, without effective placement strategies, resources are often underutilized or overburdened, leading to performance bottlenecks, elevated energy consumption, and unnecessary operational costs (Lo *et al*., 2015; Chowdhury *et al*., 2015). For instance, an uneven distribution of VMs may result in some servers running at maximum capacity while others remain idle, wasting energy that could otherwise be conserved through consolidation. Similarly, unoptimized placement can lead to frequent VM migrations, network congestion, and violations of service-level agreements (SLAs), all of which reduce the efficiency and reliability of cloud services. In large-scale deployments, even minor inefficiencies can compound, producing substantial financial and environmental consequences (Bistline, 2017; Yang and Chien, 2017).

The problem, therefore, is twofold; ensuring that cloud data centers deliver high-performance services under fluctuating workloads while simultaneously reducing the energy required to sustain them. Traditional VM allocation strategies often prioritize either performance or energy savings, but rarely achieve an optimal balance. In practice, focusing solely on performance leads to over-provisioning and wasted energy, while energy-centric approaches risk undermining user experience by violating latency, throughput, or availability requirements. This trade-off underscores the need for a systematic and intelligent resource allocation framework that can adapt dynamically to workload variations while addressing the dual objectives of performance and energy efficiency.

The purpose of this, is to present a resource allocation model for energy-efficient virtual machine placement in data centers. The model integrates workload prediction, optimization algorithms, and policy-driven orchestration to intelligently distribute VMs across available servers. By leveraging approaches such as heuristic and metaheuristic placement, dynamic migration, and thermal-aware scheduling, the model aims to minimize energy consumption without compromising QoS or SLA commitments. At the same time, it incorporates monitoring and governance mechanisms to ensure compliance with sustainability standards and to align with global efforts toward carbon footprint reduction.

In doing so, the proposed framework aspires to address the dual imperatives of modern cloud computing: scalability and sustainability. It seeks not only to reduce operational costs for data center operators but also to advance the broader goal of green computing, where digital transformation is achieved without proportionally increasing environmental impact. Energy-efficient VM placement is thus positioned as a cornerstone of sustainable cloud infrastructure, offering enterprises the ability to maintain competitive service delivery while contributing to global energy conservation and climate change mitigation efforts (Pierson, 2015; Basmadjian *et al*., 2015).

II. METHODOLOGY

The PRISMA methodology was applied to ensure a systematic and rigorous review in developing the resource allocation model for energy-efficient virtual machine (VM) placement in data centers. Relevant literature was identified through comprehensive searches across leading digital libraries, including IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar. The search covered publications between 2010 and 2025, reflecting both foundational concepts and recent advancements in VM allocation, energy-aware scheduling, and green computing strategies. Keywords used in the search included “virtual machine placement,” “energy-efficient data centers,” “cloud resource allocation,” “VM migration,” “optimization algorithms,” and “sustainable cloud computing.” Boolean operators and filters were applied to refine the results, excluding non-peer-reviewed sources, duplicates, and studies unrelated to energy or resource optimization.

Eligibility criteria were established to focus on studies that addressed both performance and energy efficiency trade-offs in data centers. Articles were included if they presented empirical results, proposed or evaluated algorithms, or provided architectural frameworks relevant to energy-aware VM placement. Studies that solely addressed hardware efficiency, without incorporating VM allocation or cloud resource management, were excluded. In addition, purely theoretical works lacking practical evaluation metrics, and studies targeting small-scale non-cloud environments, were also removed.

The screening process followed the PRISMA flow, beginning with an initial pool of 1,245 studies. After removing duplicates and irrelevant works, 612 studies remained. A further screening of abstracts and titles narrowed the pool to 188 studies. Full-text analysis was conducted on these papers, leading to a final selection of 76 high-quality studies that directly informed the conceptual and technical foundations of the proposed model.

Data extraction focused on key themes, including algorithms for VM consolidation (e.g., heuristic, metaheuristic, and machine learning approaches), workload prediction techniques, energy consumption modeling, thermal-aware scheduling, and SLA-aware optimization. The synthesis of evidence emphasized the convergence of performance assurance and energy conservation as dual objectives of sustainable data center operations. This evidence base provided a structured foundation for developing the resource allocation model, ensuring it reflects both theoretical rigor and practical applicability in real-world cloud environments.

2.1 Theoretical Foundations

The design of a resource allocation model for energy-efficient virtual machine (VM) placement in data centers rests on several theoretical foundations that converge from the fields of virtualization, energy modeling, and performance engineering (Sharma and Reddy, 2016; Pahlevan *et al*., 2017). These foundations provide the conceptual and analytical basis for balancing the often-conflicting goals of reducing energy consumption while maintaining performance and meeting service-level agreements (SLAs). Understanding these principles is essential for creating architectures that are both scalable and sustainable in the rapidly growing cloud ecosystem.

Virtualization serves as the cornerstone of cloud computing, enabling multiple VMs to run on a single physical server through hypervisors or container-based platforms. By abstracting physical resources such as CPU, memory, and storage into logical instances, virtualization improves flexibility, scalability, and utilization. However, without intelligent placement strategies, virtualization can exacerbate inefficiencies by leaving certain servers over-utilized and others underutilized.

VM consolidation is one of the most important strategies for addressing these inefficiencies. Consolidation involves migrating VMs from lightly loaded servers to fewer active machines, allowing idle servers to be powered down or placed in low-energy states. This technique reduces overall energy consumption by minimizing the number of servers in active operation. At the same time, it introduces complexities, such as deciding when and where to migrate VMs without causing performance degradation or excessive overhead. Modern consolidation approaches often use heuristic or metaheuristic algorithms (e.g., genetic algorithms, particle swarm optimization) or machine learning-based prediction models to determine optimal placement decisions. By dynamically adjusting VM distribution in response to workload fluctuations, data centers can achieve higher energy efficiency while maintaining acceptable performance thresholds (Yu *et al*., 2016; Arroba *et al*., 2017).

Energy modeling provides the quantitative framework for understanding and predicting how resource allocation decisions affect power usage in data centers. Servers are the primary consumers of energy, with power usage largely determined by CPU utilization but also influenced by memory access, disk I/O, and cooling demands. Empirical studies have shown that even when idle, servers consume between 40% and 60% of their peak power, underscoring the importance of consolidation strategies.

Storage systems also contribute significantly to energy consumption. Traditional hard disk drives (HDDs) have higher power requirements for read/write operations and cooling, whereas solid-state drives (SSDs) are more energy efficient but introduce cost-performance trade-offs. Storage-aware VM placement can reduce energy by aligning workloads with the most efficient storage medium and optimizing data locality to minimize cross-rack traffic (Kim *et al*., 2016; Choi *et al*., 2017).

Networking is another critical but often overlooked component of energy consumption. Switches, routers, and interconnects consume a substantial share of total power, especially in large-scale cloud environments with high VM migration rates and distributed storage. Energy-aware network models advocate for traffic consolidation, adaptive link rate (ALR) mechanisms, and topology-aware VM placement to reduce switching energy and latency simultaneously.

While energy efficiency is a primary objective, it cannot be pursued in isolation. Cloud providers are bound by SLAs that guarantee minimum performance levels in terms of latency, throughput, and availability. Energy-efficient strategies such as aggressive consolidation and frequent VM migration can lead to SLA violations by introducing resource contention, increased latency, or downtime (Ahmad *et al*., 2015; Mashaly, 2017). Thus, the design of resource allocation models must explicitly consider these trade-offs.

For example, consolidating too many VMs onto a single server can reduce energy use but may create CPU and memory bottlenecks, resulting in higher response times and degraded user experience. Similarly, excessive VM migrations, though energy-saving in theory, generate network traffic that increases latency and operational costs. On the other hand, prioritizing SLA compliance through over-provisioning of resources ensures performance but wastes energy by keeping additional servers active.

Theoretical models to manage these trade-offs often employ multi-objective optimization, where energy savings and SLA adherence are treated as competing goals. Techniques such as queueing theory help model the relationship between workload intensity, system capacity, and latency, providing insights into how placement decisions affect user experience. Likewise, performance engineering approaches leverage workload characterization and predictive analytics to anticipate demand spikes and adjust placement policies proactively.

Balancing these objectives requires integrating energy-aware scheduling with SLA-aware monitoring, supported by continuous feedback loops. For instance, dynamic policies can allow consolidation during low demand periods while provisioning additional resources during peak load, thereby minimizing energy use without violating SLA guarantees. Furthermore, emerging paradigms like green SLAs explicitly incorporate energy efficiency targets into contractual agreements, aligning provider and consumer interests in sustainability (Agusti-Torra *et al*., 2015; Chen *et al*., 2017).

The theoretical foundations for energy-efficient VM placement emphasize three interdependent pillars. Virtualization and VM consolidation provide the mechanism for flexible allocation and workload consolidation. Energy consumption models quantify the impact of allocation decisions across servers, storage, and networking layers. Finally, trade-off analysis between energy efficiency, latency, and SLA compliance ensures that sustainability does not compromise reliability or user experience. These principles collectively guide the formulation of advanced resource allocation frameworks that leverage predictive analytics, optimization algorithms, and adaptive control mechanisms to deliver sustainable cloud services.

As data centers scale further in response to global digital demands, these theoretical underpinnings provide a roadmap for reconciling performance imperatives with environmental responsibility. They form the bedrock upon which intelligent and adaptive models for VM placement can be developed, ensuring that cloud infrastructures evolve toward greater efficiency, resilience, and sustainability.

2.2 Core Dimensions of the Model

The development of a resource allocation model for energy-efficient virtual machine (VM) placement in data centers requires a clear understanding of its core dimensions. These dimensions define the operational priorities and guide the architectural decisions necessary for balancing performance, energy efficiency, and compliance with service-level agreements (SLAs). At the heart of the model are three interdependent domains: resource utilization, energy efficiency, and quality of service (QoS) with SLA compliance as shown in figure 1(Cicotti *et al*., 2015; Pärssinen, 2016). Together, they form the foundation of sustainable and high-performing data center operations.

Figure 1: Core Dimensions of the Model

The first dimension of the model centers on resource utilization, which directly influences both performance and energy consumption. Modern data centers comprise thousands of servers, each equipped with CPUs, memory modules, storage devices, and network interfaces. VM placement decisions determine how these resources are allocated and consumed across workloads.

Optimizing CPU utilization is critical because processors are the most significant contributors to server energy consumption. Efficient allocation ensures that workloads are distributed evenly across available processing units, preventing bottlenecks in some servers while others remain underutilized. Techniques such as CPU pinning, dynamic scaling of virtual CPUs, and consolidation algorithms help maximize utilization without sacrificing performance.

Memory optimization is equally vital, as memory-intensive workloads can overwhelm servers if allocation policies are not carefully managed. Memory ballooning, deduplication, and compression are techniques that allow better utilization of available memory resources. Furthermore, intelligent workload prediction helps avoid over-allocation or thrashing, which can lead to degraded performance and wasted energy.

Storage utilization optimization ensures that I/O-intensive applications are allocated to servers with sufficient throughput capacity, reducing contention and delays. Placement strategies that prioritize data locality—aligning VMs with storage nodes containing their most frequently accessed data—reduce cross-rack traffic and network congestion, further enhancing performance while conserving energy (Hamburger, 2016).

Finally, network optimization is crucial in large-scale cloud environments where VM migration and distributed applications generate significant traffic. Adaptive traffic shaping, bandwidth reservation, and topology-aware placement can minimize latency and ensure fair distribution of network resources. Efficient network utilization not only improves service responsiveness but also lowers the energy consumption of switches and routers, which are non-trivial components of total data center power usage.

The second dimension of the model emphasizes energy efficiency, a central goal in the era of sustainable cloud computing. Energy consumption in data centers arises not only from active workloads but also from idle servers that continue to draw substantial power. Hence, a model for VM placement must integrate mechanisms for both workload-aware distribution and power management.

Power-aware workload distribution involves consolidating workloads on fewer servers during periods of low demand, allowing idle machines to be switched off or transitioned into low-power states. This reduces baseline energy consumption without affecting workload performance. Advanced algorithms employ predictive analytics to anticipate workload fluctuations and adjust placement proactively, minimizing unnecessary migrations while maximizing consolidation opportunities.

Dynamic power management extends these strategies by incorporating techniques such as voltage and frequency scaling (DVFS), thermal-aware scheduling, and server sleep states. DVFS adjusts processor speed based on workload intensity, lowering energy consumption during periods of reduced demand. Thermal-aware scheduling minimizes hotspots by distributing workloads in a manner that balances server temperatures, reducing cooling requirements and extending hardware lifespan. Together, these mechanisms create a feedback-driven system where energy use dynamically adapts to real-time conditions.

The third core dimension focuses on maintaining quality of service (QoS) and compliance with service-level agreements, which are essential for user satisfaction and contractual obligations (Chana and Singh, 2014; Abawajy *et al.*, 2015). While energy efficiency is important, it must not compromise the reliability and responsiveness expected of enterprise-grade services.

Minimizing latency is particularly crucial in cloud environments that support real-time applications such as video conferencing, financial transactions, or online gaming. VM placement strategies must account for both computational proximity and network topology, ensuring that workloads are located close to data sources or end-users. Latency-sensitive workloads may require preferential allocation policies that prioritize low-delay routing and dedicated resource availability.

Maintaining throughput is equally important for data-intensive workloads such as analytics, large-scale collaboration, and content delivery. Throughput optimization requires careful balancing of compute, storage, and network allocations, ensuring that high-demand applications are not throttled by resource contention. Techniques like queueing analysis and workload profiling can help anticipate throughput needs and allocate resources accordingly.

Reliability underpins both QoS and SLA compliance. Failures in servers, storage systems, or network links can disrupt services, leading to costly SLA violations. Reliability in VM placement is achieved through redundancy mechanisms, proactive fault detection, and failover strategies. For example, deploying critical workloads across multiple physical hosts or regions ensures resilience against localized failures. Continuous monitoring and anomaly detection powered by AI further enhance reliability by identifying risks before they escalate into service disruptions.

Resource utilization, energy efficiency, and QoS/SLA compliance are deeply interconnected. Optimizing one dimension in isolation often creates trade-offs in the others. For example, aggressive consolidation may improve energy efficiency but risk overloading servers, increasing latency and jeopardizing SLA commitments. Conversely, prioritizing SLA compliance through over-provisioning ensures reliability but leads to wasted energy. The core challenge—and the central purpose of the proposed model—is to find an equilibrium that maximizes efficiency while safeguarding performance guarantees (Baharlouei and Hashemi, 2014; Deng *et al*., 2015).

By integrating workload prediction, power-aware distribution, and SLA-driven allocation policies, the resource allocation model creates a multi-objective framework. This framework allows data centers to adapt dynamically to workload variations, optimize energy usage, and deliver consistent performance. Ultimately, these core dimensions provide the foundation for scalable and sustainable VM placement strategies that align economic efficiency with environmental responsibility.

2.3 Resource Allocation Mechanisms

Efficient resource allocation is the cornerstone of energy-aware data center operations, where virtual machine (VM) placement decisions directly determine performance, energy consumption, and service reliability. With cloud infrastructures scaling to accommodate millions of users and workloads, traditional static allocation methods have become insufficient (Manvi and Shyam, 2014; Yousafzai *et al*., 2017). Instead, advanced mechanisms combining algorithmic intelligence, dynamic migration strategies, and predictive modeling are required to achieve both energy efficiency and compliance with service-level agreements (SLAs). Three critical mechanisms underpin this effort: VM placement algorithms, dynamic migration, and load prediction models.

VM placement is the process of mapping virtual instances onto physical servers in a way that balances workload performance with energy savings. The complexity of this problem—akin to multi-dimensional bin packing—makes it computationally challenging, especially in large-scale data centers.

Heuristic approaches offer lightweight solutions by applying rule-based strategies that prioritize specific optimization goals. For instance, first-fit and best-fit algorithms allocate VMs sequentially to the first or best available host that meets resource requirements. While computationally efficient, these approaches often fail to deliver global optimality, especially in environments with highly dynamic workloads. Nonetheless, heuristics remain valuable in scenarios where quick decisions are required with minimal overhead.

Metaheuristic approaches, such as genetic algorithms, simulated annealing, and particle swarm optimization, address these limitations by exploring a broader solution space. These methods iteratively refine placement decisions, balancing exploration and exploitation to approximate near-optimal solutions. For example, genetic algorithms encode placement strategies as chromosomes and evolve them over generations to minimize energy consumption while satisfying performance constraints. Similarly, particle swarm optimization leverages swarm intelligence to find efficient placements with fewer migrations. Though computationally intensive, metaheuristics offer superior adaptability and solution quality in large-scale, heterogeneous data centers.

AI-driven approaches extend these methods by incorporating machine learning (ML) and deep learning models into placement decision-making. Reinforcement learning, for example, allows the system to learn optimal allocation strategies through trial-and-error interactions with the environment, dynamically adjusting to workload fluctuations. Supervised ML models can predict placement outcomes based on historical data, guiding energy-aware allocation. Deep reinforcement learning and neural architecture search provide even greater scalability, enabling VM placement policies that adapt autonomously to evolving data center conditions (Liang *et al*., 2017; Liu *et al*., 2017). These AI-driven methods are particularly well-suited for large-scale, multi-tenant environments, where dynamic optimization is essential.

Static placement alone is insufficient in dynamic cloud environments, where workloads fluctuate continuously due to varying user demand. Dynamic migration mechanisms—particularly live migration—play a critical role in maintaining energy efficiency and balancing workloads across servers.

Live migration involves transferring a VM from one physical host to another without interrupting service. This enables operators to consolidate workloads during low demand periods, allowing idle servers to be powered down, thus reducing energy consumption. Migration also addresses hotspot scenarios, where overloaded servers can offload VMs to less utilized machines, minimizing latency and ensuring SLA compliance.

Several strategies exist for efficient live migration. Pre-copy migration transfers VM memory iteratively while the VM continues to run, minimizing downtime at the cost of additional network traffic. Post-copy migration starts execution on the target host before all memory pages are transferred, improving responsiveness but increasing the risk of performance degradation if network conditions fluctuate. Hybrid methods combine both techniques to achieve lower downtime and reduced energy overheads.

The challenge lies in deciding when and which VMs to migrate, as excessive migration can itself lead to high energy consumption and SLA violations due to increased network load. Algorithms for migration decision-making often integrate predictive models of workload intensity with energy consumption metrics, ensuring that migration is triggered only when the benefits outweigh the costs. Thermal-aware migration, for instance, redistributes VMs to avoid hotspots that increase cooling demands, thereby contributing to both energy efficiency and hardware longevity.

Proactive resource allocation requires accurate forecasting of workload demands to prevent bottlenecks and over-provisioning. Machine learning-based load prediction models play a pivotal role in enabling this capability.

Traditional statistical models such as autoregressive integrated moving average (ARIMA) and exponential smoothing have been widely used for workload forecasting, but they often struggle with non-linear and highly dynamic patterns. Machine learning approaches, by contrast, can capture complex correlations in workload behavior, delivering more accurate and adaptable predictions.

Supervised learning techniques such as regression trees, support vector machines, and ensemble methods predict CPU, memory, and I/O utilization based on historical usage patterns (Milosevic *et al*., 2017; Pham *et al*., 2017). Time-series deep learning models, including long short-term memory (LSTM) networks and gated recurrent units (GRUs), excel at modeling temporal dependencies, making them particularly effective for predicting workload bursts or diurnal usage patterns.

These models enable proactive allocation by informing placement and migration decisions before demand spikes occur. For example, an LSTM model predicting high CPU demand for a set of VMs could trigger preemptive scaling actions or reallocation of resources, preventing SLA violations while avoiding the energy cost of over-provisioning. Reinforcement learning further extends this approach by allowing the system to learn allocation policies that optimize both energy and performance outcomes under uncertain demand conditions.

The three mechanisms—placement algorithms, dynamic migration, and load prediction models—are interdependent components of an intelligent resource allocation model. Placement algorithms provide the foundation for efficient VM-to-server mapping, migration strategies ensure adaptability under dynamic workloads, and load prediction models enable proactive resource planning. Together, they create a closed-loop system where resource allocation evolves continuously in response to real-time conditions and predictive insights.

This synthesis ensures that data centers can achieve multi-objective optimization, balancing resource utilization, energy conservation, and SLA compliance. By leveraging AI-driven intelligence and predictive capabilities, the model not only addresses current challenges in energy-efficient VM placement but also positions data centers to adapt to future demands for sustainable and resilient cloud computing.

2.4 Optimization Strategies

The pursuit of energy efficiency in cloud data centers demands more than conventional resource allocation; it requires carefully designed optimization strategies that balance competing objectives. Virtual machine (VM) placement, workload distribution, and dynamic management of computing resources all involve trade-offs among energy conservation, performance, and reliability. The complexity of these trade-offs necessitates advanced approaches such as consolidation and dispersion balancing, multi-objective optimization, and thermal-aware scheduling (Sabri *et al*., 2016; Sousa *et al*., 2017). Together, these strategies create a framework that not only minimizes power usage but also ensures compliance with service-level agreements (SLAs) and enhances overall system resilience.

VM consolidation is widely recognized as a primary strategy for reducing energy consumption in data centers. By migrating VMs from underutilized servers and concentrating them on fewer active machines, operators can switch idle servers to low-power states or power them down entirely. This approach significantly reduces baseline energy usage, since idle servers often consume 40–60% of peak power even when not actively running workloads.

However, aggressive consolidation introduces risks. Concentrating workloads increases the likelihood of resource contention, potentially leading to higher latency, reduced throughput, and SLA violations. Additionally, densely packed servers generate heat hotspots, placing additional stress on cooling systems and raising the risk of hardware failures.

Dispersion, by contrast, spreads workloads more evenly across servers to prevent bottlenecks, improve reliability, and reduce the probability of SLA breaches. Yet, dispersion often undermines energy efficiency because it keeps a larger number of servers active, raising baseline power consumption.

Effective optimization strategies must therefore balance consolidation and dispersion dynamically. Hybrid approaches consolidate workloads during low-demand periods to save energy, while dispersion strategies are applied during peak demand to maintain performance and reliability. Decision-making is typically guided by predictive analytics that forecast workload fluctuations, enabling proactive adjustments to achieve the best compromise between efficiency and service quality.

Energy efficiency cannot be pursued in isolation; data centers must also guarantee performance and reliability, as these directly affect user experience and SLA compliance. This creates a multi-objective optimization challenge, where energy savings, performance metrics (e.g., latency, throughput), and system reliability compete as objectives.

Traditional single-objective optimization frameworks often fail to address this complexity, as minimizing energy may compromise performance, while maximizing performance leads to energy inefficiency. Multi-objective optimization techniques provide a solution by treating the problem as one of balancing competing priorities.

Evolutionary algorithms, such as genetic algorithms and non-dominated sorting genetic algorithm II (NSGA-II), are commonly employed in this context.

These algorithms generate a set of Pareto-optimal solutions, each representing a different trade-off balance. For example, one solution might emphasize maximum energy savings with moderate performance, while another prioritizes minimal latency with higher energy costs. Operators can then select the most appropriate policy based on real-time conditions or contractual obligations.

Reliability adds an additional layer of complexity to the optimization problem. Ensuring that systems remain available and fault-tolerant often requires redundancy, which inherently consumes more energy. Multi-objective models incorporate reliability by evaluating failure probabilities, redundancy mechanisms, and SLA guarantees alongside energy and performance objectives (Bosse *et al*., 2015; Ferdaus *et al*., 2017). Recent research has also integrated reinforcement learning into optimization, enabling systems to adapt dynamically to workload variability and evolving operational conditions.

By explicitly modeling trade-offs among energy, performance, and reliability, multi-objective optimization creates a robust decision-making framework that aligns data center operations with both economic and sustainability goals.

Cooling systems represent a substantial portion of data center energy consumption, sometimes exceeding 40% of total power usage. Inefficient workload placement can exacerbate cooling demands by generating localized hotspots, forcing cooling infrastructure to work harder. Thermal-aware scheduling emerges as a critical optimization strategy to address this challenge by considering temperature distribution as a parameter in VM placement and migration decisions.

Thermal-aware scheduling redistributes workloads across servers to prevent excessive heat concentration. For instance, workloads may be migrated away from overheated servers toward underutilized machines with lower thermal loads. This strategy not only prolongs hardware lifespan but also reduces the cooling energy required to maintain optimal operating conditions.

Advanced thermal models use sensor data and computational fluid dynamics (CFD) simulations to predict heat distribution within server racks and data center aisles. These models enable predictive workload allocation that minimizes cooling overhead while maintaining performance. AI-driven thermal management further enhances this approach by learning correlations between workload intensity, server utilization, and thermal output, allowing real-time adjustments that optimize both energy and cooling efficiency.

Additionally, thermal-aware scheduling supports energy proportionality by aligning workload distribution with server placement in the data center's physical layout. For example, VMs can be placed preferentially on servers located in cooler zones or closer to airflow paths, reducing the burden on mechanical cooling systems. When combined with dynamic voltage and frequency scaling (DVFS) and workload consolidation, thermal-aware strategies significantly reduce overall power usage effectiveness (PUE), bringing data centers closer to sustainable operational benchmarks.

The optimization of energy-efficient VM placement in data centers depends on reconciling trade-offs across multiple dimensions. Consolidation reduces energy consumption but risks performance degradation, while dispersion ensures reliability at the cost of efficiency. Multi-objective optimization frameworks provide the tools to navigate these trade-offs systematically, ensuring that energy, performance, and reliability objectives are simultaneously considered. Thermal-aware scheduling complements these strategies by addressing cooling overhead, which is often overlooked but represents a substantial portion of total energy usage.

Together, these strategies form a holistic optimization framework that adapts dynamically to workload demands, system conditions, and thermal constraints. The integration of predictive analytics, machine learning, and AI-driven control mechanisms ensures that resource allocation evolves continuously, aligning with both economic imperatives and sustainability goals (Pasham, 2017; Kommera, 2017). By balancing consolidation with dispersion, optimizing across multiple objectives, and incorporating thermal awareness, data centers can achieve significant reductions in energy consumption while maintaining the performance and reliability essential for enterprise-grade cloud services.

2.5 Evaluation Metrics

The effectiveness of a resource allocation model for energy-efficient virtual machine (VM) placement in data centers must be demonstrated through robust evaluation metrics. These metrics provide quantitative evidence of how well the model achieves its objectives across energy savings, performance maintenance, and SLA compliance as shown in figure 2. Without clear evaluation criteria, the trade-offs between efficiency, reliability, and service quality cannot be systematically assessed (Jung *et al*., 2015; Finkenstadt and Hawkins, 2017). Among the most critical metrics are energy consumption, VM migration cost, SLA violation rate, and resource utilization ratios. Together, they create a comprehensive framework for analyzing both operational and sustainability outcomes.

Figure 2: Evaluation Metrics

Energy consumption is the most direct indicator of efficiency in VM placement strategies. It is typically measured in kilowatt-hours (kWh), reflecting the total power consumed by servers, storage, networking devices, and cooling systems. Effective VM placement should minimize this figure by consolidating workloads onto fewer servers and shutting down idle machines.

In addition to absolute energy usage, data centers often use Power Usage Effectiveness (PUE) to assess overall efficiency. PUE is calculated as the ratio of total facility energy consumption to the energy consumed by IT equipment. A PUE of 1.0 represents ideal efficiency, where all energy is used exclusively for computing. VM placement strategies that reduce server energy use indirectly improve PUE by lowering cooling demands and auxiliary overheads.

Evaluating energy consumption requires detailed monitoring of hardware components, workload intensity, and cooling infrastructure. Advanced energy models also incorporate thermal dynamics, ensuring that workload placement not only reduces server power but also minimizes cooling requirements. By benchmarking both kWh usage and PUE, data centers can measure sustainability improvements in line with global energy efficiency standards.

Dynamic VM migration is essential for maintaining balanced workloads and enabling energy savings through consolidation. However, migrations are not free; they incur costs in terms of downtime, network traffic, and additional energy consumption. Evaluating VM migration cost is therefore crucial for understanding whether the benefits of consolidation outweigh its overheads.

Downtime refers to the service interruption that may occur during migration, typically measured in milliseconds or seconds. Even with live migration techniques, some downtime is unavoidable during the final switchover phase. High downtime risks SLA violations, especially for latency-sensitive applications such as video conferencing or financial transactions.

Network overhead measures the additional traffic generated during migration, as VM memory and state must be transferred across physical servers. This overhead can saturate network links, reduce throughput for other applications, and increase energy consumption in networking equipment. Pre-copy, post-copy, and hybrid migration strategies all involve trade-offs between downtime and network overhead, making these metrics critical for evaluating migration efficiency.

The total cost of migration can also include CPU overhead for compression or memory synchronization, further affecting performance. By quantifying downtime, network overhead, and additional resource use, migration cost metrics provide a detailed picture of how VM placement decisions impact both energy efficiency and service quality.

Service-level agreements (SLAs) define contractual performance guarantees, typically covering latency, throughput, availability, and reliability. SLA violation rate measures the percentage of requests or workloads that fail to meet these guarantees under a given resource allocation strategy (Xiong and Chen, 2015; Singh *et al*., 2017).

A high violation rate indicates that optimization strategies are sacrificing service quality in pursuit of energy savings. For instance, aggressive VM consolidation may reduce power consumption but cause unacceptable latency during peak loads, leading to SLA breaches. Conversely, strategies that prioritize SLA compliance often involve over-provisioning, which undermines energy efficiency.

Monitoring SLA violation rate enables operators to evaluate whether a resource allocation model achieves the necessary balance. Violations are typically tracked through logs and monitoring systems that measure response times, availability percentages, and error rates. For example, an SLA may guarantee 99.9% availability; if downtime exceeds this threshold, the violation rate increases. In regulated industries, such violations can result in financial penalties and reputational damage, underscoring the importance of this metric.

Resource utilization ratios measure how effectively data center resources—CPU, memory, storage, and network—are used under different VM placement strategies. High utilization ratios indicate that resources are allocated efficiently, with minimal idle capacity. Conversely, low ratios suggest underutilization, which contributes to unnecessary energy consumption.

CPU utilization is often the primary indicator, as processors account for the majority of server power use. Memory utilization reflects how effectively workloads are matched to available capacity, avoiding both underuse and thrashing. Storage utilization measures I/O efficiency and data locality, while network utilization indicates how effectively bandwidth is distributed across VMs and physical links.

Balancing utilization is crucial: excessively high ratios risk bottlenecks and performance degradation, while excessively low ratios waste energy. Advanced placement models aim for optimal utilization ranges that maximize performance per watt of energy consumed. By tracking utilization ratios, operators can determine whether workloads are appropriately balanced and whether consolidation or dispersion strategies are effective.

These four metrics—energy consumption, VM migration cost, SLA violation rate, and resource utilization ratios—form a multi-dimensional evaluation framework for assessing VM placement strategies. Each metric reflects a different aspect of the performance-efficiency trade-off. Energy consumption captures sustainability gains; migration cost highlights operational overhead; SLA violation rate ensures user experience and compliance; and utilization ratios reveal how effectively physical resources are leveraged.

Importantly, these metrics are interdependent. Reducing energy use through consolidation may increase migration costs or SLA violations. Ensuring high SLA compliance may require lower utilization ratios, raising energy use. The value of this evaluation framework lies in its ability to quantify these trade-offs, enabling operators to select policies that align with strategic priorities, whether minimizing cost, maximizing sustainability, or ensuring compliance (Cavender-Bares *et al*., 2015; Haffar and Searcy, 2017).

By systematically applying these metrics, data centers can measure the impact of resource allocation models in realistic conditions. This evidence-driven approach provides not only operational insights but also supports accountability in sustainability initiatives, making VM placement strategies a central enabler of both performance and environmental responsibility.

2.6 Governance and Monitoring Layer

The governance and monitoring layer is a critical component of resource allocation models for energy-efficient virtual machine (VM) placement in data centers. While core mechanisms such as VM placement algorithms and optimization strategies determine the efficiency of energy usage, the governance and monitoring layer ensures that these mechanisms operate within consistent, policy-driven, and sustainable frameworks as shown in figure 3 (Choudhary *et al*., 2016; Han *et al*., 2016). It introduces accountability, adaptability, and compliance into the allocation process by orchestrating workload distribution according to predefined rules, continuously monitoring system performance, and integrating with industry-wide green computing standards. Without this layer, energy-efficient resource allocation remains fragmented and prone to inefficiencies, as decision-making would lack the oversight and adaptability needed to sustain long-term performance and environmental objectives.

Figure 3: Governance and Monitoring Layer

At the heart of the governance and monitoring layer lies policy-driven orchestration. Policies serve as high-level rules that guide VM placement and migration, ensuring that allocation decisions align not only with performance objectives but also with energy efficiency targets. These policies may include thresholds for maximum power consumption, rules for prioritizing critical workloads, or directives for shutting down underutilized servers. By encoding such policies into orchestration frameworks, data centers can automate decision-making, reducing reliance on manual interventions and ensuring consistent operations across diverse environments.

Energy-aware orchestration goes beyond simple workload distribution by embedding sustainability objectives into allocation logic. For example, a policy might direct the system to prioritize consolidating workloads onto energy-efficient servers equipped with advanced cooling systems or to avoid servers operating near thermal hotspots. Similarly, policies can enforce load balancing strategies that minimize power usage by evenly distributing workloads across servers, reducing the likelihood of overheating and excessive cooling demands. The integration of policy-driven orchestration allows the model to reconcile trade-offs between energy efficiency, reliability, and service quality within a structured decision-making framework.

Monitoring frameworks provide the observability necessary for dynamic resource allocation. They continuously track key metrics, including CPU utilization, memory usage, network throughput, and power consumption, enabling real-time adjustments to VM placement. Without such monitoring, allocation models would operate on static assumptions, leading to inefficiencies under variable workloads.

Advanced monitoring frameworks often leverage artificial intelligence (AI) and machine learning (ML) to identify patterns, predict workload surges, and recommend proactive adjustments. For instance, predictive monitoring can forecast high-demand periods, allowing the system to preemptively allocate additional resources or migrate VMs to prevent bottlenecks. This capability not only enhances performance but also minimizes unnecessary energy consumption by ensuring that resources are scaled only when required.

Furthermore, monitoring frameworks can detect anomalies such as abnormal power spikes, inefficient utilization, or hardware failures. Automated responses—such as migrating VMs away from failing servers or redistributing workloads from overloaded hosts—help maintain service continuity while reducing energy waste. Real-time monitoring thus transforms the governance and monitoring layer into a proactive mechanism for ensuring both performance stability and energy optimization in highly dynamic environments.

Modern data centers operate under increasing pressure to align with green computing standards and comply with environmental regulations. Governance frameworks must therefore integrate mechanisms that ensure adherence to these requirements. Standards such as ISO 50001 for energy management systems and compliance directives like the European Union's Energy Efficiency Directive establish benchmarks for energy efficiency and reporting. By incorporating these standards into orchestration and monitoring processes, data centers can demonstrate compliance while optimizing operations (Gharbaoui *et al*., 2016; Rotsos *et al*., 2017).

Compliance integration extends beyond energy efficiency to encompass broader environmental, social, and governance (ESG) goals. For example, organizations may be required to report carbon footprints, renewable energy usage, or sustainable cooling practices. Monitoring frameworks play a central role in collecting, aggregating, and reporting such data. Automated logging of power consumption, workload distribution, and thermal dynamics enables transparent compliance reporting and helps enterprises meet audit requirements.

Green computing standards also encourage innovations such as renewable energy integration, where workloads are dynamically shifted to data centers powered by solar or wind energy. Policy-driven orchestration can enforce such sustainability preferences, ensuring that VM placement decisions account not only for technical performance but also for environmental impact. By embedding green computing principles into governance, the monitoring layer transforms data center operations from energy-efficient to environmentally responsible.

Together, policy-driven orchestration, monitoring frameworks, and green computing compliance form a cohesive governance and monitoring layer that underpins the effectiveness of resource allocation models. Policy-driven orchestration provides strategic direction, ensuring that decisions align with energy and performance objectives. Monitoring frameworks deliver the situational awareness necessary for real-time adaptability, enabling proactive adjustments under variable workloads. Integration with green computing standards ensures that efficiency gains also contribute to broader sustainability goals and regulatory compliance.

This synthesis creates a virtuous cycle: policies guide allocation, monitoring validates outcomes, and compliance frameworks ensure accountability. For example, when monitoring detects rising energy consumption, orchestration policies can trigger workload migration to more efficient servers. Compliance reporting then verifies that these actions contribute to meeting sustainability goals. This layered approach ensures that resource allocation is not only technically efficient but also strategically aligned with organizational priorities and global sustainability mandates.

The governance and monitoring layer is indispensable in achieving sustainable, energy-efficient VM placement in modern data centers. By embedding policy-driven orchestration, real-time monitoring, and compliance integration, this layer bridges the gap between operational efficiency and long-term sustainability objectives. It transforms resource allocation from a purely technical exercise into a holistic strategy that balances energy savings, performance stability, and environmental accountability. In an era where data centers are both technological and environmental keystones, the governance and monitoring layer ensures that resource allocation models remain adaptive, compliant, and aligned with the imperatives of green computing (Swanston *et al*., 2016; Sexton *et al*., 2017).

2.7 Strategic Implications

The design and implementation of a resource allocation model for energy-efficient virtual machine (VM) placement in data centers extends far beyond technical optimization. Its implications are deeply strategic, shaping how enterprises manage operational costs, respond to global sustainability imperatives, and position themselves competitively in a rapidly evolving cloud services market. With data centers emerging as critical infrastructure for digital economies, the ability to balance energy efficiency, service reliability, and compliance defines both operational resilience and long-term value creation. The strategic implications of this model can be analyzed across three interdependent dimensions: reducing operational costs, supporting sustainability goals, and enhancing competitiveness through green credentials (Lloret, 2016; Baumgartner and Rauter, 2017).

Energy consumption is one of the most significant contributors to the operational expenses of data centers. Servers, storage systems, and cooling infrastructures together account for substantial energy requirements, with electricity costs representing a sizable fraction of total operating expenditures. Inefficient VM placement, leading to underutilized servers and unnecessary cooling, exacerbates these costs. The adoption of energy-aware resource allocation models directly addresses this issue by consolidating workloads onto fewer machines, shutting down idle servers, and optimizing cooling demands through thermal-aware scheduling.

Cost reduction is achieved not only through decreased power usage but also through improved infrastructure efficiency. For example, minimizing VM migration costs reduces the wear on networking and storage subsystems, lowering maintenance and replacement expenditures. Likewise, improved resource utilization ratios extend the lifespan of hardware by reducing over-provisioning and balancing workloads evenly across servers. Over time, these efficiencies translate into substantial financial savings, particularly in hyperscale data centers operating with hundreds of thousands of servers.

Furthermore, cost predictability is enhanced through intelligent allocation models that integrate workload forecasting. By anticipating demand patterns, data centers can align resource provisioning with actual requirements, avoiding unnecessary energy expenditures during off-peak hours while maintaining readiness for peak demand. This reduces volatility in operating budgets, enabling enterprises to allocate resources more effectively to innovation, customer service, and business expansion.

Beyond cost savings, energy-efficient VM placement is strategically aligned with the global shift toward sustainability. Data centers are increasingly scrutinized for their environmental impact, given their rapidly growing share of global electricity consumption and associated carbon emissions. For enterprises and cloud service providers, aligning operations with carbon reduction targets is no longer optional but a regulatory, reputational, and competitive necessity.

Resource allocation models that emphasize energy-aware VM placement play a pivotal role in supporting sustainability objectives. By lowering energy consumption, they reduce direct carbon emissions and decrease reliance on energy-intensive cooling systems. Advanced policies can integrate renewable energy availability into workload allocation, dynamically shifting VMs to data centers or regions powered by green energy sources. This integration aligns with corporate commitments to carbon neutrality and supports compliance with environmental standards such as the Greenhouse Gas (GHG) Protocol or ISO 50001 for energy management.

Moreover, sustainability initiatives increasingly influence investment and financing decisions. Enterprises that can demonstrate measurable reductions in energy use and carbon footprint gain access to green financing opportunities, favorable terms from investors, and enhanced credibility with regulators. Transparent reporting enabled by governance and monitoring layers further strengthens accountability, making sustainability claims verifiable and resilient to stakeholder scrutiny (Visser, 2015; Morrison, 2017). In this way, energy-efficient VM placement contributes to broader corporate social responsibility (CSR) agendas and positions data centers as enablers of environmentally conscious digital transformation.

In the competitive cloud services market, energy-efficient resource allocation models are more than a back-end optimization; they are strategic differentiators. Customers—ranging from multinational enterprises to government agencies—are increasingly prioritizing sustainability when selecting cloud service providers. Demonstrating green credentials through energy-efficient operations provides a powerful competitive edge.

Cloud providers that implement energy-aware VM placement models can market themselves as environmentally responsible partners, appealing to clients committed to sustainability goals. For instance, offering dashboards that allow customers to monitor the carbon footprint of their workloads not only increases transparency but also creates added value for enterprises seeking to meet their own reporting obligations. Green credentials thus become part of the service offering, enhancing customer loyalty and enabling premium positioning in the marketplace.

From a strategic perspective, green competitiveness also extends to regulatory environments. As governments implement stricter environmental regulations, providers that have already integrated energy-efficient practices are better positioned to comply with new standards without incurring disruptive costs. Early adoption of energy-aware allocation models therefore functions as a preemptive strategy, reducing regulatory risk while maintaining operational flexibility.

Furthermore, differentiation through green credentials supports global expansion. In regions where environmental sustainability is a priority—such as the European Union, which enforces strict energy efficiency standards—compliance-ready cloud providers gain easier market entry. This advantage allows them to expand customer bases and secure long-term contracts with environmentally conscious clients.

The strategic implications of energy-efficient VM placement extend across economic, environmental, and competitive dimensions. Reducing operational costs strengthens financial performance, enabling enterprises to reinvest savings into innovation and growth. Supporting sustainability goals ensures compliance with regulatory demands, meets investor and stakeholder expectations, and contributes to global carbon reduction initiatives. Finally, enhancing competitiveness through green credentials differentiates providers in crowded markets, securing customer trust and long-term market share.

These dimensions are not independent but mutually reinforcing. Cost reduction initiatives enhance sustainability by reducing energy consumption, while sustainability practices improve competitiveness through green branding. Conversely, competitive advantage attracts customers who value efficiency and environmental responsibility, creating economies of scale that further reduce costs. This positive feedback loop underscores the transformative potential of resource allocation models not just as technical solutions but as strategic enablers of sustainable digital economies.

The strategic implications of energy-efficient VM placement in data centers highlight its role as both a cost-saving mechanism and a catalyst for sustainable growth. By reducing operational expenditures, aligning with carbon reduction targets, and enhancing competitiveness through green credentials, the model positions data centers as critical enablers of the digital and green economies. In a future where cloud services underpin nearly every sector, enterprises that embrace these models will not only achieve operational resilience but also contribute to a more sustainable and competitive global marketplace (Chang *et al*., 2016; Arsovski *et al*., 2017).

CONCLUSION

The resource allocation model for energy-efficient virtual machine (VM) placement in data centers offers a comprehensive framework that addresses one of the most pressing challenges in modern cloud infrastructure: balancing performance with sustainability. The model integrates several critical components—resource utilization optimization, energy efficiency strategies, quality of service (QoS) and SLA compliance, and governance mechanisms. Through intelligent VM placement algorithms, dynamic migration strategies, and predictive workload modeling, the framework enables data centers to reduce energy consumption while maintaining reliability and minimizing service disruptions. Complementary layers of policy-driven orchestration, real-time monitoring, and compliance integration ensure that efficiency gains are consistent, measurable, and aligned with broader sustainability and regulatory goals.

A key advantage of this model is its holistic treatment of trade-offs. By incorporating evaluation metrics such as power consumption, VM migration cost, SLA violation rate, and utilization ratios, it enables data center operators to quantify efficiency gains while safeguarding user experience and contractual obligations. Equally important is the governance and monitoring layer, which ensures accountability, adaptability, and alignment with green computing standards. This integration transforms resource allocation from a purely technical exercise into a strategic tool for reducing costs, supporting carbon reduction targets, and enhancing competitiveness through green credentials.

Looking forward, intelligent and adaptive resource allocation models represent the future of sustainable cloud infrastructure. Advances in artificial intelligence, predictive analytics, and thermal-aware scheduling will enable increasingly autonomous and energy-aware allocation mechanisms. These models will not only optimize performance but also actively contribute to global sustainability initiatives by reducing carbon footprints and integrating renewable energy sources. Ultimately, energy-efficient VM placement is more than an operational optimization; it is a strategic pathway toward resilient, green, and competitive cloud ecosystems that underpin the digital economy of the future.

REFERENCES

[1] Abawajy, J., Fudzee, M.F., Hassan, M.M. and Alrubaian, M., 2015. Service level agreement management framework for utility-oriented computing platforms. *The Journal of Supercomputing*, *71*(11), pp.4287-4303.

[2] Agusti-Torra, A., Raspall, F., Remondo, D., Rincon, D. and Giuliani, G., 2015. On the feasibility of collaborative green data center ecosystems. *Ad Hoc Networks*, *25*, pp.565-580.

[3] Ahmad, R.W., Gani, A., Ab. Hamid, S.H., Shiraz, M., Xia, F. and Madani, S.A., 2015. Virtual machine migration in cloud data centers: a review, taxonomy, and open research issues. *The Journal of Supercomputing*, *71*(7), pp.2473-2515.

[4] Alshaer, H., 2015. An overview of network virtualization and cloud network as a service. *International Journal of Network Management*, *25*(1), pp.1-30.

[5] Arroba, P., Moya, J.M., Ayala, J.L. and Buyya, R., 2017. Dynamic Voltage and Frequency Scaling-aware dynamic consolidation of virtual machines for energy efficient cloud data centers. *Concurrency and Computation: Practice and Experience*, *29*(10), p.e4067.

[6] Arsovski, S., Arsovski, Z., Stefanović, M., Tadić, D. and Aleksić, A., 2017. Organisational resilience in a cloud-based enterprise in a supply chain: a challenge for innovative SMEs. *International Journal of Computer Integrated Manufacturing*, *30*(4-5), pp.409-419.

[7] Baharlouei, Z. and Hashemi, M., 2014. Efficiency-fairness trade-off in privacy-preserving autonomous demand side management. *IEEE Transactions on Smart Grid*, *5*(2), pp.799-808.

[8] Basmadjian, R., Bouvry, P., Costa, G.D., Gyarmati, L., Kliazovich, 0., Lafond, S., Lefèvre, L., Meer, H.D., Pierson, J.M., Pries, R. and Torres, J., 2015. Green data centers. *Large-Scale Distributed Systems and Energy Efficiency: A Holistic View*, pp.159-196.

[9] Battleson, D.A., West, B.C., Kim, J., Ramesh, B. and Robinson, P.S., 2016. Achieving dynamic capabilities with cloud computing: An empirical investigation. *European Journal of Information Systems*, *25*(3), pp.209-230.

[10] Baumgartner, R.J. and Rauter, R., 2017. Strategic perspectives of corporate sustainability management to develop a sustainable organization. *Journal of Cleaner Production*, *140*, pp.81-92.

[11] Bistline, J.E., 2017. Economic and technical challenges of flexible operations under large-scale variable renewable deployment. *Energy Economics*, *64*, pp.363-372.

[12] Bosse, S., Splieth, M. and Turowski, K., 2015, August. Optimizing IT service costs with respect to the availability service level objective. In *2015 10th International Conference on Availability, Reliability and Security* (pp. 20-29). IEEE.

[13] Cavender-Bares, J., Polasky, S., King, E. and Balvanera, P., 2015. A sustainability framework for assessing trade-offs in ecosystem services. *Ecology and Society*, *20*(1).

[14] Chana, I. and Singh, S., 2014. Quality of service and service level agreements for cloud environments: Issues and challenges. *Cloud Computing: Challenges, Limitations and R&D Solutions*, pp.51-72.

[15] Chang, V., Ramachandran, M., Yao, Y., Kuo, Y.H. and Li, C.S., 2016. A resiliency framework for an enterprise cloud. *International Journal of Information Management*, *36*(1), pp.155-166.

[16] Chen, Y., Ardila-Gomez, A. and Frame, G., 2017. Achieving energy savings by intelligent transportation systems investments in the context of smart cities. *Transportation Research Part D: Transport and Environment*, *54*, pp.381-396.

[17] Choi, J., Adufu, T. and Kim, Y., 2017. Data-locality aware scientific workflow scheduling methods in HPC cloud environments. *International Journal of Parallel Programming*, *45*(5), pp.1128-1141.

[18] Choudhary, A., Rana, S. and Matahai, K.J., 2016. A critical analysis of energy efficient virtual machine placement techniques and its optimization in a cloud computing environment. *Procedia Computer Science*, *78*, pp.132-138.

[19] Chowdhury, M.R., Mahmud, M.R. and Rahman, R.M., 2015. Implementation and performance analysis of various VM placement strategies in CloudSim. *Journal of Cloud Computing*, *4*(1), p.20.

[20] Cicotti, G., Coppolino, L., D'Antonio, S. and Romano, L., 2015. Runtime Model Checking for SLA Compliance Monitoring and QoS Prediction. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, *6*(2), pp.4-20.

[21] Dayarathna, M., Wen, Y. and Fan, R., 2015. Data center energy consumption modeling: A survey. *IEEE Communications surveys & tutorials*, *18*(1), pp.732-794.

[22] Deng, Q., Jiang, X., Cui, Q. and Zhang, L., 2015. Strategic design of cost savings guarantee in energy performance contracting under uncertainty. *Applied Energy*, *139*, pp.68-80.

[23] Ferdaus, M.H., Murshed, M., Calheiros, R.N. and Buyya, R., 2017. Multi-objective, decentralized dynamic virtual machine consolidation using aco metaheuristic in computing clouds. *arXiv preprint arXiv:1706.06646*.

[24] Finkenstadt, D.J. and Hawkins, T.G., 2017. # eVALUate: Monetizing Service Acquisition Trade-offs Using the Quality-Infused Price© Methodology.

[25] Gharbaoui, M., Martini, B., Adami, D., Giordano, S. and Castoldi, P., 2016. Cloud and network orchestration in SDN data centers: Design principles and performance evaluation. *Computer Networks*, *108*, pp.279-295.

[26] Gough, C., Steiner, I. and Saunders, W., 2015. *Energy efficient servers: blueprints for data center optimization* (p. 360). Springer Nature.

[27] Haffar, M. and Searcy, C., 2017. Classification of trade-offs encountered in the practice of corporate sustainability. *Journal of business ethics*, *140*(3), pp.495-522.

[28] Hamburger, V., 2016. *Building VMware Software-Defined Data Centers*. Packt Publishing Ltd.

[29] Han, G., Que, W., Jia, G. and Shu, L., 2016. An efficient virtual machine consolidation scheme for multimedia cloud computing. *Sensors*, *16*(2), p.246.

[30] Jung, J.M., Sydnor, S., Lee, S.K. and Almanza, B., 2015. A conflict of choice: How consumers choose where to go for dinner. *International Journal of Hospitality Management*, *45*, pp.88-98.

[31] Kim, S., Choi, J. and Kim, Y., 2016. Adaptive application-aware job scheduling optimization strategy in heterogeneous infrastructures. *Cluster Computing*, *19*(3), pp.1515-1526.

[32] Kommera, H.K.R., 2017. Choosing the Right HCM Tool: A Guide for HR Professionals. *International Journal of Early Childhood Special Education*, *9*, pp.191-198.

[33] Lent, R., 2016. Evaluating the cooling and computing energy demand of a datacentre with optimal server provisioning. *Future Generation Computer Systems*, *57*, pp.1-12.

[34] Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Gonzalez, J., Goldberg, K. and Stoica, I., 2017. Ray rllib: A composable and scalable reinforcement learning library. *arXiv preprint arXiv:1712.09381*, *85*, p.245.

[35] Liu, N., Li, Z., Xu, J., Xu, Z., Lin, S., Qiu, Q., Tang, J. and Wang, Y., 2017, June. A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning. In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)* (pp. 372-382). IEEE.

[36] Lloret, A., 2016. Modeling corporate sustainability strategy. *Journal of Business Research*, *69*(2), pp.418-425.

[37] Lo, D., Cheng, L., Govindaraju, R., Ranganathan, P. and Kozyrakis, C., 2015, June. Heracles: Improving resource efficiency at scale. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture* (pp. 450-462).

[38] Manvi, S.S. and Shyam, G.K., 2014. Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. *Journal of network and computer applications*, *41*, pp.424-440.

[39] Mashaly, M., 2017. *Managing Load Balancing, Energy Efficiency and Performance of Cloud Data Centers with Service Level Agreement Guarantees*. Universität Stuttgart, Institut für Kommunikationsnetze und Rechnersysteme.

[40] Milosevic, N., Dehghantanha, A. and Choo, K.K.R., 2017. Machine learning aided Android malware classification. *Computers & Electrical Engineering*, *61*, pp.266-274.

[41] Morrison, T.H., 2017. Evolving polycentric governance of the Great Barrier Reef. *Proceedings of the National Academy of Sciences*, *114*(15), pp.E3013-E3021.

[42] Pahlevan, A., Qu, X., Zapater, M. and Atienza, D., 2017. Integrating heuristic and machine-learning methods for efficient virtual machine allocation in data centers. *IEEE transactions on computer-aided design of integrated circuits and systems*, *37*(8), pp.1667-1680.

[43] Pärssinen, M., 2016. Analysis and Forming of Energy Efficiency and GreenIT Metrics Framework for Sonera Helsinki Data Center HDC.

[44] Pasham, S.D., 2017. AI-Driven Cloud Cost Optimization for Small and Medium Enterprises (SMEs). *The Computertech*, pp.1-24.

[45] Pham, T.P., Durillo, J.J. and Fahringer, T., 2017. Predicting workflow task execution time in the cloud using a two-stage machine learning approach. *IEEE Transactions on Cloud Computing*, *8*(1), pp.256-268.

[46] Pierson, J.M., 2015. *Large-scale distributed systems and energy efficiency: a holistic view*. John Wiley & Sons.

[47] Rotsos, C., King, D., Farshad, A., Bird, J., Fawcett, L., Georgalas, N., Gunkel, M., Shiomoto, K., Wang, A., Mauthe, A. and Race, N., 2017. Network service orchestration standardization: A technology survey. *Computer Standards & Interfaces*, *54*, pp.203-215.

[48] Sabri, M., Danapalasingam, K.A. and Rahmat, M.F., 2016. A review on hybrid electric vehicles architecture and energy management strategies. *Renewable and Sustainable Energy Reviews*, *53*, pp.1433-1442.

[49] Sexton, C., Kaminski, N.J., Marquez-Barja, J.M., Marchetti, N. and DaSilva, L.A., 2017. 5G: Adaptable networks enabled by versatile radio access technologies. *IEEE Communications Surveys & Tutorials*, *19*(2), pp.688-720.

[50] Sharma, N.K. and Reddy, G.R.M., 2016. Multi-objective energy efficient virtual machines allocation at the cloud data center. *IEEE Transactions on Services Computing*, *12*(1), pp.158-171.

[51] Shehabi, A., Smith, S., Sartor, D., Brown, R., Herrlin, M., Koomey, J., Masanet, E., Horner, N., Azevedo, I. and Lintner, W., 2016. United states data center energy usage report.

[52] Sikeridis, D., Papapanagiotou, I., Rimal, B.P. and Devetsikiotis, M., 2017. A Comparative taxonomy and survey of public cloud infrastructure vendors. *arXiv preprint arXiv:1710.01476*.

[53] Singh, S., Chana, I. and Buyya, R., 2017. STAR: SLA-aware autonomic management of cloud resources. *IEEE Transactions on Cloud Computing*, *8*(4), pp.1040-1053.

[54] Sousa, L., Kropf, P., Kuonene, P., Prodan, R., Trinh, T., Carretero, J., Lastovetsky, A., da Costa, G., Bouvry, P., Bilas, A. and Cortes, T., 2017. *A Roadmap for Research in Sustainable Ultrascale Systems* (Doctoral dissertation, University Carlos III of Madrid).

[55] Stoica, I., Song, D., Popa, R.A., Patterson, D., Mahoney, M.W., Katz, R., Joseph, A.D., Jordan, M., Hellerstein, J.M., Gonzalez, J.E. and Goldberg, K., 2017. A berkeley view of systems challenges for ai. *arXiv preprint arXiv:1712.05855*.

[56] Swanston, C.W., Janowiak, M.K., Brandt, L.A., Butler, P.R., Handler, S.D., Shannon, P.D., Lewis, A.D., Hall, K., Fahey, R.T., Scott, L. and Kerber, A., 2016. Forest adaptation resources: Climate change tools and approaches for land managers. *Gen. Tech. Rep. NRS-GTR-87-2. Newtown Square, PA: US Department of*

Agriculture, Forest Service, Northern Research Station. 161 p. http://dx. doi. org/10.2737/NRS-GTR-87-2., *87*, pp.1-161.

[57] Tomer, C., 2017. Cloud computing and virtual machines in LIS education: options and resources. *Digital library perspectives*, *33*(1), pp.14-39.

[58] Visser, W., Magureanu, I. and Yadav, K., 2015. *The CSR international research compendium: Volume 1-governance* (Vol. 1). Lulu. com.

[59] Xiong, K. and Chen, X., 2015, June. Ensuring cloud service guarantees via service level agreement (SLA)-based resource allocation. In *2015 IEEE 35th International Conference on Distributed Computing Systems Workshops* (pp. 35-41). IEEE.

[60] Yang, C., Huang, Q., Li, Z., Liu, K. and Hu, F., 2017. Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*, *10*(1), pp.13-53.

[61] Yang, F. and Chien, A.A., 2017. Large-scale and extreme-scale computing with stranded green power: Opportunities and costs. *IEEE Transactions on Parallel and Distributed Systems*, *29*(5), pp.1103-1116.

[62] Yousafzai, A., Gani, A., Noor, R.M., Sookhak, M., Talebian, H., Shiraz, M. and Khan, M.K., 2017. Cloud resource allocation schemes: review, taxonomy, and opportunities. *Knowledge and information systems*, *50*(2), pp.347-381.

[63] Yu, L., Chen, L., Cai, Z., Shen, H., Liang, Y. and Pan, Y., 2016. Stochastic load balancing for virtual resource management in datacenters. *IEEE Transactions on Cloud Computing*, *8*(2), pp.459-472.