International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Review Study on Energy Consumption in Cloud Computing

Oğuzhan Şereflişan[1], Havelsan[2], Murat Koyuncu[3]

[1]Department of Software Engineering, Atilim University, Ankara, Turkey
[2]Ankara, Turkey
[3]Department of Information Systems Engineering, Atilim University, Ankara, Turkey

Abstract:
Cloud computing has become a fundamental technology for a wide range of computing services, yet its increasing energy demands present substantial environmental and economic challenges. With the rapid growth of Cloud services and applications, increasing number of researches have been focused on energy saving. The need to reduce energy costs is a constant challenge of cloud providers and data centers. This paper offers an extensive review of the issues surrounding energy consumption in cloud computing, with a focus on algorithms associated with the situational awareness, consolidation, allocation, placement/migration, and scheduling of virtual machines and containers. We conduct a critical analysis of studies from 2018 to 2023, comparing various methodologies aimed at achieving energy efficiency without sacrificing performance. This review delineates current trends, identifies gaps in existing research, and proposes directions for future investigations. Our study emphasizes the necessity of cultivating sustainable practices in cloud computing and provides valuable insights into the practical implementation of energy-efficient solutions in cloud environments.

Keywords: cloud computing, energy efficiency, virtual machines, containers

1. Introduction

Data centers are at the heart of modern software technology. They play a critical role in expanding the business world's ability to do much more with much less, both in terms of physical space, money and the time required to create and maintain data.

As shown in Figure 1, the energy consumption source of the data center can be roughly divided into two parts [23]: energy consumption of IT facilities and infrastructure energy consumption. IT facilities refer to information technology facilities such as server systems (including servers and storage) and network systems, while the supporting infrastructure mainly includes cooling systems and power supply systems (including power supply systems and lighting systems).

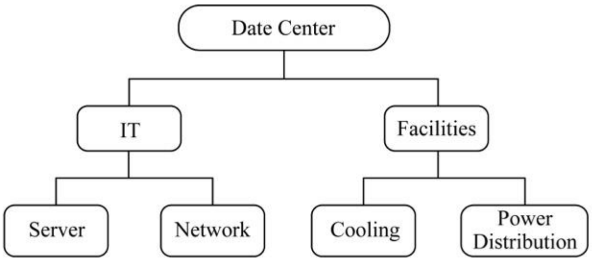

Figure 1 Data center energy consumption source decomposition

Early versions of data centers are like isolated resource areas with physical servers, cooling systems and networking equipment used for enterprise backup systems or data storage such as silos. Technological developments are giving rise to new-generation data centers. One of the innovations is cloud computing, in which computing services such as programs, storage space, expert services, video games, films and music are made available on demand via an online service. Mobile development has also triggered a high demand for online services and energy [23].

Cloud computing is founded on the principles of Service-Oriented Architecture (SOA) inherent in Web 2.0, along with the virtualization of both hardware and software resources. Virtualization technology provides more opportunities to data centers about giving adequate resources for computing and consuming less energy. Virtualization also offers data centers the opportunity to process more with the same infrastructure. Thus, virtualization and research on virtualization became popular. There are several other ways to reduce energy consumption related to information processing resources such as storage systems, cache management systems, processor architecture, schedulers, etc. But the popularity of virtualization motivates us to do this overview study.

Rapid development of cloud computing technologies and applications caused cloud data centers have grown exponentially in recent years. One of the major problems with current cloud data centers is their huge energy consumption that makes management of energy consumption one of the hottest research topics.

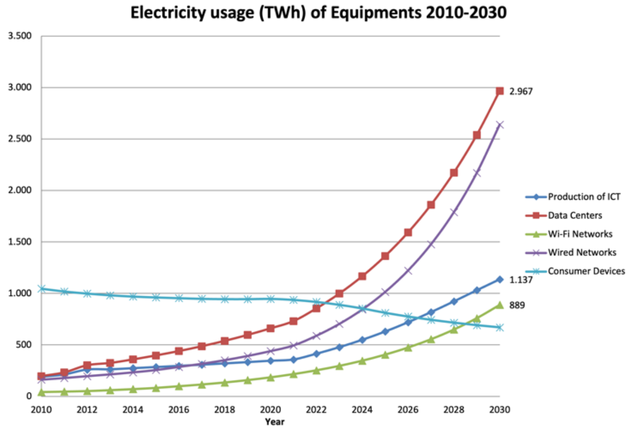

Figure 2 Energy consumption forecast [2]

Figure 2 presents an "expected case" projection by Andres Andrae, who is a specialist in sustainable Information and Communication Technology (ICT). In his best scenario, datacenter energy consumption is growing 15 times higher during 2010 to 2030. Also energy consumption in wired networking is growing nearly the same.

Data centers currently consume an estimated 200 terawatt hours (TWh) annually. This amount exceeds the national energy consumption of several countries, including Iran. It accounts for half of the global electricity used for transportation [1].

Figure 3 shows the energy supply including only wholesale colocation facilities in North America. In Europe, Latin America, and Asia-Pacific, total inventory includes both wholesale and retail colocation facilities. Inventory continues to climb, with Northern Virginia leading CBRE's global rankings. It has 2,132 MW (2.1 GW) of supply, increasing 19.5% year-over-year from Q1 2022 to Q1 2023. Data center supply grew year-over-year in Frankfurt, London, Amsterdam and Paris (FLAP) as providers work to meet the strong demand across most top European markets. There is 672 MW of inventory in Latin America as of Q1 2023, primarily across Brazil, Mexico, Chile and Colombia. The region has experienced significant growth over the past three years, with supply doubling since Q1 2020. Brazil has grown the fastest, with inventory up 127% from 2020 to 2022. It is also the largest, with around 67% of the region's inventory. Tokyo, Sydney and Singapore each now contain over a half-GW of live power capacity. Sydney's inventory jumped 30% year-over-year. Lack of power availability is a key emerging challenge facing operators in some Asia-Pacific markets.

According to the report for European Commission [5], data centers in London consumes 386 MWs, in Frankfurt 199 MWs, in Amsterdam 166 MWs and in Paris 129 MWs. Figure 4 shows energy consumption detail in data centers in Germany.

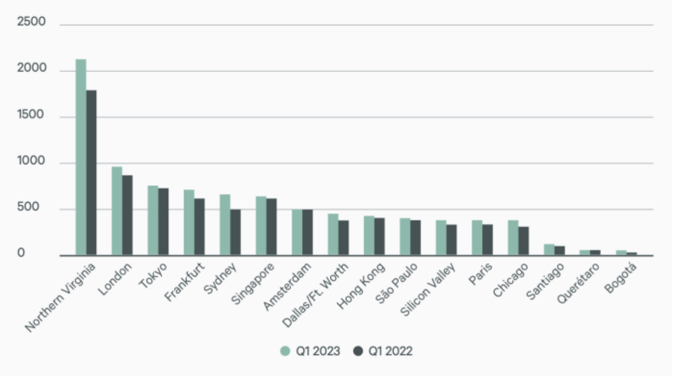

Figure 3 Inventory graph for energy consumptions (MW) [3]

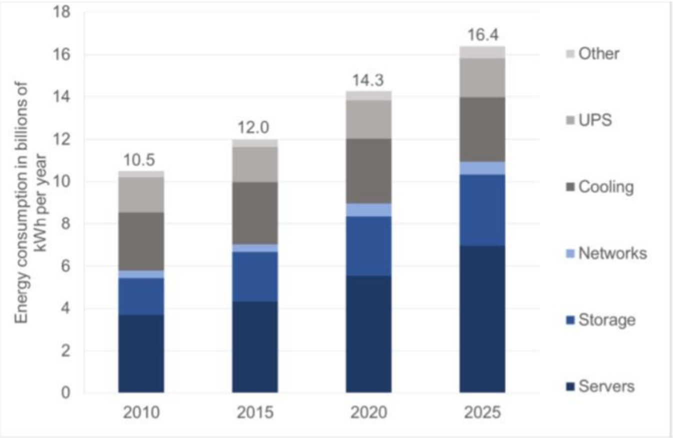

Figure 4 The energy consumption of servers and data centers in Germany, from 2010 to 2015, and including forecast up to 2025[5]

In general, all data centers have similar energy consumption values. When analyzing the reports defined above, we examined the power consumption of the servers, as this accounts for the largest share in the graph in Figure 4. Research into servers or computing power is a very good way of reducing the energy consumption of data centers.

In recent years, technological developments in servers have led to lower energy requirements than expected. Use of energy-efficient processors, memory and storage devices in cloud infrastructure provides companies to reduce overall energy consumption. In addition to such improvements that can be made in terms of hardware on the servers, software developments are also being made. In our study, we do not focus on the hardware equipment in the data center.

Both VMs and containers benefit from energy-efficient hardware. Especially developments in server and processing technology in a way of virtualization, make cloud computing more attractive for researchers.

Virtual machines (VMs) and containers are increasingly being used in data center environments to improve resource utilization, reduce hardware costs, and simplify application deployment. However, the use of VMs and containers also leads to increased energy consumption, which can have significant financial and environmental impacts.

Energy consumption in managing VMs and containers can be attributed to several factors, including inactive VMs and containers, over-provisioned resources, inefficient workload placement, and cooling and power distribution systems. Various algorithms and techniques have been proposed to optimize resource allocation, scheduling and power management in VM and container management. If the energy efficiency of VMs and containers is taken into account in management, data centers can achieve significant energy savings and reduce their carbon footprints.

Similar to VM management, container management can also benefit from energy-efficient resource allocation, migration and scheduling. Containerization has emerged as an alternative to virtualization, offering similar benefits but with a lower overhead in terms of resource usage and energy consumption.

Container management is a complex task involving multiple factors such as resource allocation, workload scheduling, placement or migration and situational power monitoring. In recent years, containerization has become increasingly popular in academic studies.

In light of the information provided in Figure 4, our study has been conducted with a focus on server energy consumption in cloud computing environments. This study specifically concentrates on evaluating algorithms related to virtualization and containerization within cloud computing. The focus is on how these algorithms impact CPU usage, which is the primary consumer of power in a server. These algorithms are thoroughly assessed from various perspectives, encompassing aspects such as situational awareness, scheduling, consolidation, allocation, and placement/migration.

Cloud computing and virtualization methods are explained in Section II and the virtualization management system is provided in Section III. The used algorithms for situational awareness, scheduling, consolidation, allocation, and placement/migration of virtual machines and/or containers are examined in Section IV.

1. BREAKTHROUGH IN CLOUD COMPUTING

Existing cloud data centers are fully virtualized for service consolidation and energy reduction. Additionally, motivations for adopting virtualization encompass enhanced flexibility, dynamic resource allocation, and improved resource utilization. As a result, this paradigm has gained considerable attractiveness, leading to the development of various solutions over the years. Broadly, these solutions can be categorized into two types: hypervisor-based virtualization and Operating System (OS)-based virtualization methods.

Hypervisor-based methods are generally used and become traditional technique for many cloud providers. For example, Amazon Web Services (AWS) and Rackspace use the XEN Hypervisor [7] which has gained popularity because of its early open source inclusion in the Linux kernel, and is one of the most mature virtualization solutions available [8]. The Kernel-based Virtual Machine (KVM) [9], a relatively new open source hypervisor-based system, has gained momentum and popularity in recent years [10]. KVM has become a natural choice for Linux VMs because it is included in the upstream Linux kernel. Also there are VMWare [16] , VirtualBox [20] and XEN [21] in hypervisor-based virtualization.

Hypervisor-free OS-based virtualization such as container-based virtualization provides multiple isolated user spaces with only one running kernel instance, while hypervisor-based systems operate at the hardware abstraction level. Because this virtualization offers higher performance without the need to invoke virtual machines and guest operating systems, hypervisor-free OS-based virtualization systems are widely used by successful cloud providers such as Google. In recent years, most of cloud providers provide hypervisor-free OS-based virtualization choice to consumers. This choice offers new possibilities for easy provisioning and fast deployment environment. Hypervisor-free OS-based virtualization serves as a lightweight alternative to traditional hypervisor-based virtualizations. For example, Google has stated that each week they start over billion containers across all of its services [11]. Several OS-based systems have been released, including Linux Container (LXC) [12], Docker [13], BSD Jails [14], and Windows Containers [15]. Docker is the most widely used.

In virtualization there are two ways for presentation of a software interface to virtual machines, which are full-virtualization and paravirtualization. Paravirtualization presents a modified interface to virtual machines and achieves better performance [22] because applications are not running on an extra operating system. Full-virtualization presents higher compatibility by completely emulating the underlying hardware and this makes virtual machine like emulating a computer. This feature makes full-virtualization have more overhead than paravirtualization in because of running an extra full operating system on host operating system.

2.1. VIRTUALIZATION AND CONTAINERIZATION

Figure 4 clearly illustrates that within the realm of cloud computing, servers are responsible for the largest share of energy consumption. This underlines the significant impact that reducing server energy consumption can have on enhancing overall energy efficiency in cloud environments. It is important to acknowledge the extensive research and development efforts that have been dedicated to managing energy consumption, particularly focusing on Central Processing Unit (CPU) management and network management.

Virtualization allows multiple virtual machines (VMs) to share the same physical resources, which leads to better utilization of hardware and reduces the overall energy consumption. The rapid expansion in both number and scale of cloud data centers globally is a direct consequence of the increasing number of cloud computing users. This expansion has led to a more pronounced problem of energy consumption. The issue is multifaceted: it escalates operational costs and diminishes the return on investment for cloud service providers, while also contributing significantly to carbon dioxide emissions, thereby exacerbating global warming. Consequently, this problem has garnered considerable attention from both cloud service providers and governmental bodies.

Containerization represents a methodology in software development and deployment that enables applications to operate reliably and consistently across various computing environments. This approach involves encapsulating an application along with all its dependencies into a unified package, known as a container. This container can then be executed on any system equipped with the requisite containerization technology.

A container provides a complete runtime environment for an application, including the operating system, libraries, and any other dependencies that the application needs to run. This allows developers to create applications that can be easily moved between different environments, such as development, testing, and production, without needing to modify the application code or its dependencies.

Containerization technology is typically based on the use of container engines such as Docker, LXC, etc. , which allow containers to be built, shipped, and run on a wide range of platforms, including on-premises data centers, public cloud services, and edge computing devices.

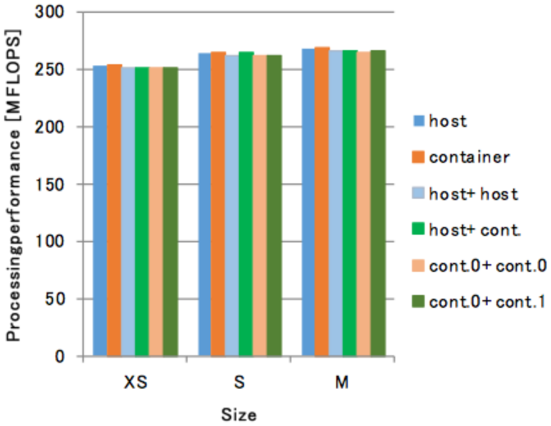

Figure 5 CPU processing performance (Host and container) [18]

Virtual Machines (VMs) and containers are two widely utilized virtualization technologies, implemented at the hardware and Operating System (OS) levels, respectively. Consequently, containers can be deployed on VMs, allowing them to complement each other and function more effectively in tandem.

In Figure 5 you can see the performance of containers and hybrid use. Containers are lighter than VMs because they share the kernel of the host operating system, resulting in lower overhead and potentially lower energy consumption. They are also more portable and can be started and stopped more quickly, making them suitable for dynamic workloads.

2.2. COMPARISON OF VIRTUALIZATION AND CONTAINERIZATION

Virtualization and containerization are two different approaches to manage resources in data centers. Virtualization uses hypervisors to create multiple virtual machines, while containerization uses a single operating system kernel to create multiple containers. Both approaches offer benefits in terms of resource utilization and energy consumption. However, containerization has several advantages over virtualization in terms of energy consumption.

Containerization has a lower overhead in terms of resource usage and energy consumption than virtualization. Containers use the same operating system kernel, which reduces memory footprint and CPU utilization compared to virtual machines. This results in a lower energy consumption with containerization compared to virtualization.

Containerization also allows more efficient workload scheduling and resource allocation compared to virtualization. Containers can be quickly started and stopped, which allows for more flexible workload scheduling. Resource allocation in containers is also more efficient as they share the same operating system kernel, which allows more efficient resource utilization

Virtual machine (VM) management refers to the process of creating, deploying, and managing virtual machines in a data center environment. VMs are software-based representations of physical hardware, which can be used to run multiple operating systems and applications on a single physical server. VM management involves the allocation of resources, such as CPU, memory, and storage, to VMs and the monitoring and optimization of VM performance[24].

Container management bears similarities to VM management, but it employs a more lightweight and efficient approach for packaging and deploying applications, as opposed to using a full virtual machine. The process of container management encompasses the creation, deployment, and administration of containers, along with resource allocation and performance monitoring. By encapsulating an application and all its components within a container, it abstracts the underlying infrastructure and operating systems on which the application is deployed [25].

2. VIRTUALIZATION MANAGEMENT SYSTEM INTERNAL WORKFLOW

In recent years, several approaches have been proposed to optimize energy consumption in VM and container management. These approaches include dynamic workload consolidation, energy management and resource allocation algorithms. Dynamic workload consolidation involves moving VMs and containers between servers to consolidate workloads and power down unused servers. Power management is about reducing the power consumption of unused servers and components. Resource allocation algorithms aim to allocate resources efficiently by predicting the resource requirements of VMs and containers and allocating resources accordingly.

There are various algorithms that are used by virtualization managers in cloud environments. To categorize these algorithms, we need to determine the levels.

Cloud providers make agreements with their customers to define the service level, the so-called Service Level Agreements (SLA). These agreements define the requirements and priorities of the customer requests. These definitions are also used for processing customer requests for VM and container management. The virtualization managers prioritize the incoming tasks before processing the requests, and the requirements specified in the contract are also taken into account in this prioritization.

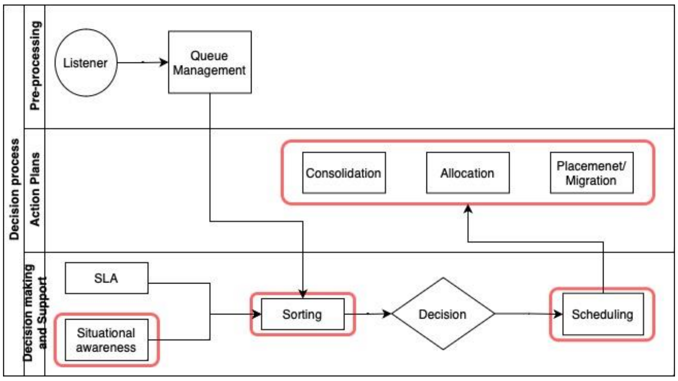

Fig.6 Workflow of a virtualization manager system decision stages for virtualization objects (VMS)

The virtualization management system comprises many subsystems and decides on the management of virtualization objects within the framework of 3 stages as seen in Figure 6. These stages are as follows: *Pre-processing*, *decision making and support* and *action plans*.

Under the heading of *pre-processing*, listener systems and queuing systems can be positioned to recognize future requests and requirements. Listener systems, as the name suggests, are always on standby and listen to identify potential requests. Requests and tasks that arrive after this stage are forwarded to the sorting systems to be sorted. Immediate metric measurements and prioritization information from the SLA, which are necessary for the ranking systems to do their job, are provided by the supporting systems. The integration of queuing and sorting processes can be efficiently achieved through the utilization of support systems. These systems are designed to handle complex tasks by organizing and prioritizing incoming data or requests, thereby enhancing overall operational efficiency.

In the decision-making and support process, performance metrics are generated based on the current situation from the virtualization system and the requirements stipulated in the service level agreements signed with customers. These metrics are then forwarded to the subsequent stage, which is the Sorting process, within the same operational workflow. During *sorting*, incoming requests are evaluated using SLA data and performance metrics, sorted using the specified algorithms and sent to the scheduler.

After the sorting process is completed, a decision action is processed by the management system. Decisions can be;

- Creating and/or placement of a virtualization component (VM or container)
- Consolidation of virtualization component
- Allocation of virtualization component
- Migration of virtualization component

The decision action triggers the scheduling process to perform the selected action in the selected time. These workflow elements contain a variety of orders for the execution of certain tasks.

The algorithms used in pre-processing process are used in decision-making and supporting and action plans. Artificial intelligence can be used, for example, in situational awareness or in the decision-making process.

Most studies focus on sorting, scheduling, consolidation, allocation and placement/migration processes.

The remainder of this section is dedicated to providing an explanation of Virtual Machine System (VMS) processes. Subsequently, in Section IV, academic articles that have been thoroughly researched and presented are provided.

a. ***Situational awareness***: Cloud providers regularly monitor and analyze resource usage to identify inefficiencies and optimize resource allocation. Monitoring tools for VMs, such as VMware vSphere and Microsoft System Center, and for containers, such as Prometheus and Grafana, can provide insights into VM and container performance and resource usage. This information can be used to optimize VM allocation, ensuring that resources are used efficiently and energy consumption is minimized.

b. ***Scheduling:*** Scheduling is a process for determining the right time to perform a certain action. We can either plan the timing ourselves or let AI determine it based on external parameters. In addition, the load distribution can be planned in this phase according to the incoming workload.

c. ***Consolidation:*** Consolidating workloads onto fewer servers is an effective strategy for maximizing resource utilization and minimizing the number of idle servers, which continue to consume energy despite being underutilized. In the context of VMs, server consolidation can be achieved by deploying multiple Virtual Machines on a single physical host. This approach reduces the necessity for numerous physical servers to host applications, thereby leading to decreased energy consumption. Regarding containers, their inherently lightweight nature allows for even higher consolidation ratios. Operating multiple containers on the same host can result in enhanced resource utilization and further reductions in energy consumption.

d. ***Allocation***: Allocation in virtual machine (VM) management is a process that involves distributing virtual components across physical systems. A key method for reducing energy consumption in this context is through effective resource allocation and scheduling. This process entails dynamically assigning resources, like CPU and memory, to VMs, taking into account their specific workload and performance needs. To optimize resource allocation and scheduling in VM management, various algorithms have been developed and proposed.

e. ***Placement/Migration*****:** Placement is the task of first placement of a virtual component to the proper physical/virtual environment. Migration is the task of moving a virtual component to another physical or virtual environment. Especially migration process is a costly process to handle.

4. LITERATURE REVIEW

Optimizing energy consumption when managing VMs and containers requires the use of various algorithms that enable efficient resource allocation, workload management, and optimization of cooling and power distribution systems. In this section, we explore and discuss several main algorithms that recent research has proposed, focusing on optimizing energy consumption in the management of Virtual Machines (VMs) and containers.

In our research, we conducted an analysis of sources from Scopus, IEEE, MDPI, Wiley, Springer, Semantic Scholar, Google Academic, ACM, and various cloud computing-related conferences. The selection criteria for articles or papers were as follows:

a. Relevance to cloud computing, explicitly excluding those containing keywords such as 'ship', 'heat', 'solar', and 'package'.
b. Publication date between 2018 and 2023.
c. Inclusion of keywords related to 'Virtual' or 'Container'.
d. Mention of 'power usage' or 'energy consumption'.
e. Exclusion of publications focusing on 'network', 'edge', 'IoT', '5G', and '6G'.

The rationale behind these criteria included:

a. To capture the trends in publications over the last five years that are pertinent to energy efficiency in cloud computing.
b. To distinguish our focus from maritime and transportation publications.
c. To concentrate on algorithms that are directly related to energy efficiency in virtual machines and containers, with a specific emphasis on CPU usage, which is our primary area of focus.
d. To segregate our research from publications primarily concerned with network, edge computing, and hardware-related topics.

Our mapping study resulted in a knowledge base of current research approaches, methods, techniques, best practices and experiences used in cloud computing, with a particular attention to virtualization management system processes as situational awareness, scheduling, consolidation, allocation and placement/migration.

4.1. SITUATIONAL AWARENESS

In this process, system is regularly monitored and analyzed in terms of resource usage to identify inefficiencies and optimize resource allocation.

Monitoring and situational awareness (SA) are crucial aspects of ensuring the health, security, and performance of cloud computing environments. Recent years have seen significant advancements in algorithms designed to collect, analyze, and interpret data for effective cloud monitoring and awareness. By leveraging the advancements in machine learning and other technologies, these algorithms are becoming increasingly sophisticated and effective in ensuring the health, security, and performance of cloud environments. Here’s specific monitoring and awareness areas:

- Security-aware Monitoring: This area focuses on monitoring for security threats and vulnerabilities in cloud environments. Papers like [47] discuss techniques for anomaly detection and intrusion prevention in cloud systems.
- Performance Monitoring: This area focuses on monitoring cloud resources and applications to ensure optimal performance. Papers like [28] address resource utilization and performance bottlenecks in cloud environments.
- Cost Optimization Monitoring: This area focuses on monitoring cloud resource usage and costs to optimize spending. Papers like [26] discuss techniques for predicting resource demands and optimizing resource allocation for cost efficiency.

Table 1 is a curated review of relevant papers published in the past five years. The table shows focus, algorithms and key features of situational awareness/monitoring studies between 2018-2023. A study [29] proposes a cloud-based framework for space situational awareness (SSA), leveraging cloud computing's scalability and flexibility for tracking and monitoring space objects. It highlights the benefits of cloud-based SSA for efficient data storage, processing, and analysis. Another study [44] proposes a reinforcement learning-based approach for resource management in cloud data centers. It demonstrates the effectiveness of RL (Reinforcement Learning) in optimizing resource allocation for performance and energy efficiency.

Paper	Focus	Approach	Key Features
[35]	Power Models	Power profiling technique	VM and server power profiling relationship
[6]	Prediction of resource allocation	Machine Learning based Prediction	Prediction based resource allocation, Improved resource allocation
[17]	Dynamic Voltage Scaling	Cooling monitoring	Distribute the load
[47]	General SA	Framework & Techniques	Design principles, data collection/analysis methods, evaluation approaches
[29]	General SA & Cloud	Cloud-based SSA framework	Scalability, flexibility for space object monitoring
[28]	Resource Optimization & Anomaly Detection	Deep learning anomaly detection	Proactive resource management, cost reduction
[26]	Resource Management & Prediction	Deep learning resource prediction	Improved resource allocation, proactive management
[4]	Anomaly Detection & Edge Computing	Federated learning for anomaly detection	Distributed anomaly detection in cloud-edge environments
[44]	Resource Management & Data Centers	Reinforcement learning resource allocation	Performance & energy efficiency optimization

[53]	Performance prediction	Machine Learning based Prediction	Compute-intensive workloads, memory-intensive workloads and I/O- intensive workloads

Table 1. Situational awareness/monitoring studies between 2018-2023 in respect of energy consumption

Predicting the power consumption attributable to a specific virtual component presents a significant challenge. In study [35], two power models have been developed that correlate the workload of a Virtual Machine (VM) with its proportion of power consumption. These models are based on the power profiling of a server. The proposed models demonstrate an improvement in accuracy, approximately by 3%, compared to other existing models in the field.

To tackle the issue of inefficient resource utilization, existing approaches primarily concentrate on Virtual Machine (VM) allocation and migration. This strategy typically results in optimization at the Physical Machine (PM) level. However, other methods employ horizontal auto-scaling, which may not be an effective solution, particularly in the context of Infrastructure as a Service (IaaS) public cloud environments. The study proposes an approach that involves customizing the size of user Virtual Machines (VMs) to align with the resource requirements of their application workloads. This approach is based on an analysis of real backend traces collected from a VM operating in a production data center. The approach detailed in [6] involves allocating fixed-size resources to a Virtual Machine (VM) that are specifically designed to meet the application workload's demands. In cases where the demands exceed this predetermined resource allocation, the strategy utilizes vertical VM auto-scaling to manage the excess requirements effectively. This method aims to reduce energy consumption by Physical Machines (PMs) through enhanced resource utilization. Experimental results, derived from a simulation conducted on CloudSim Plus and utilizing GWA-T-13 Materna real backend traces, demonstrate that efficient resource utilization can lead to a reduction in data center energy consumption by approximately 40-52%.

The research referenced in [17] introduces two Dynamic Voltage and Frequency Scaling (DVFS)-enabled host selection algorithms for Virtual Machine (VM) placement in a cluster, specifically the Carbon and Power-Efficient Optimal Frequency (C-PEF) algorithm and the Carbon-Aware First-Fit Optimal Frequency (C-FFF) algorithm. These algorithms are primarily designed to achieve two key objectives: to evenly distribute the load across servers and to dynamically adjust the cooling load in response to the current workload.

A notable aspect of these algorithms is their cluster selection strategy, which is based on both static and dynamic Power Usage Effectiveness (PUE) values, as well as the Carbon Footprint Rate (CFR). Furthermore, the research extends the cluster selection approach to encompass non-DVFS host selection policies. This extension includes the development of the Carbon- and Power-Efficient (C-PE) algorithm, Carbon-Aware First-Fit (C-FF) algorithm, and Carbon-Aware First-Fit Least-Empty (C-FFLE) algorithm. The findings from this research reveal that the C-FFF algorithm is particularly effective, achieving a 2% greater power reduction compared to the C-PEF and C-PE algorithms. This positions C-FFF as a power-efficient solution for reducing carbon dioxide emissions, while maintaining the same Quality of Service (QoS) as its counterparts. Additionally, it does so with lower computational overheads, further emphasizing its efficiency.

4.2. SCHEDULING

The dynamic and distributed nature of cloud computing necessitates efficient scheduling algorithms to optimize resource utilization, job completion times, and overall system performance.

Table 2 shows focus, algorithms and key features of scheduling studies between 2018-2023. The study [64] proposes an ant colony optimization-based algorithm that schedules tasks across heterogeneous cloud resources while minimizing energy consumption. Another study [34] leverages reinforcement learning to schedule workflow tasks in the cloud, considering deadline and budget constraints. The study [61] addresses fairness concerns in multi-tenant cloud environments by proposing a priority queue-based scheduling approach that guarantees a minimum QoS for all tenants. The paper [67] explores dynamic voltage scaling for scheduling tasks with approximation tolerance in cloud datacenters, achieving energy savings without compromising accuracy. The study [68] investigates cost-efficient scheduling of containerized workloads in the cloud by considering resource costs and execution time. Another study [63] proposes a priority-aware scheduling algorithm for serverless functions in edge computing environments, ensuring timely execution of critical tasks. The paper [65] combines reinforcement learning and heuristic techniques for scheduling tasks and allocating resources in cloud environments, leading to enhanced performance. The study [66] proposes a federated learning framework for scheduling tasks in dynamic cloud environments, enabling distributed and collaborative learning among scheduling agents. The study [69] implements deep learning practices for resource prediction.

Paper	Focus	Approach	Key Features
[60]	Energy-aware	Inverted Ant colony optimization, Capuchin Search Algorithm	Heterogeneous cloud resources scheduling and migration and decision making framework
[41]	Scheduling	Deep Reinforcement Learning	Future directions
[37]	VM Scheduling	Genetic algorithm	Minimize resource usage and energy consumption
[56]	Container scheduling	Genetic algorithm	Resource usage and energy consumption
[64]	Energy-aware	Ant colony optimization	Minimizes energy and migration overhead through metaheuristic approach
[34]	Deadline & Budget-constrained	Reinforcement learning	Schedules workflow tasks considering deadlines and budget restrictions
[61]	Fairness-aware & Multi-tenant	Priority queues	Guarantees minimum QoS for tenants while scheduling across queues
[67]	Dynamic Voltage Scaling & Approximate Computing	DVS for approximation tolerance	Saves energy by scaling voltage for tasks with acceptable approximation

[68]	Cost-aware & Containerized workloads	Cost-aware container scheduling	Optimizes execution time and resource costs for containerized workloads
[63]	Priority-aware & Edge computing	Priority-aware scheduling	Ensures timely execution of critical tasks in edge computing environments
[65]	Hybrid	Hybrid reinforcement learning & heuristics	Optimizes task scheduling and resource allocation with combined approach
[66]	Federated learning	Federated learning for dynamic clouds	Enables collaborative learning among scheduling agents for dynamic resource management
[69]	Deep learning	Deep learning for resource prediction	Predicts resource demands for improved scheduling and utilization

Table 2. Scheduling studies between 2018-2023 in respect of energy consumption

In a study [60], researchers propose an ant colony optimization-based algorithm that schedules tasks across heterogeneous cloud resources while minimizing energy consumption. The study introduces a novel approach for migrating Virtual Machines (VMs) utilizing a capuchin search algorithm (CapSA). This proposed method aims to leverage the strengths of both migration and scheduling. It does so by employing a hybrid approach that combines multi-objective CapSA and inverted ant colony optimization (IACO) algorithms. The approach involves selecting an optimal algorithm for subsequent tasks by using a decision-making framework that considers the specific conditions of the received tasks. Compared to previous methods, the proposed approach demonstrates superior performance in terms of energy consumption (EC), execution time (ET), and load balancing, showing improvements in the range of 15–20%.

In study [61], researchers introduced a time and energy-aware two-phase scheduling algorithm, named Best Heuristic Scheduling (BHS), specifically designed for scheduling Directed Acyclic Graphs (DAG) on processors in cloud data centers. This algorithm operates in two distinct phases: firstly, it sorts, and then it identifies the best performing action. The findings from this study indicate that the BHS algorithm achieves 19.71% more energy savings compared to the Multiheuristic Resource Allocation (MHRA) algorithm. Additionally, the makespan, which is the total duration required to complete execution, is reduced by 56.12% in heterogeneous environments, demonstrating the algorithm's effectiveness in optimizing both energy efficiency and execution time.

In study [34], researchers developed two workflow scheduling algorithms that leverage the structural properties of workflows. The first algorithm, named Structure-based Multi-objective Workflow Scheduling with an Optimal instance type (SMWSO), introduces a novel method for determining both the optimal instance type and the optimal number of Virtual Machines (VMs) to be provisioned. Additionally, in the Structure-based Multi-objective Workflow Scheduling with Heterogeneous instance types (SMWSH), the algorithm is adapted to include the use of heterogeneous VMs, thereby demonstrating its effectiveness in heterogeneous environments. Simulation results from this study indicate that the proposed algorithms achieve better energy efficiency in 80% of the tested workflow/workload scenarios. They also show a significant energy saving, exceeding 50%, when compared to a recent state-of-the-art algorithm, underscoring their potential for substantial energy conservation in cloud computing environments.

In study [41], a survey is conducted on resource scheduling methods, with a particular emphasis on Deep Reinforcement Learning (DRL)-based scheduling approaches in cloud computing. The study comprehensively reviews the application of DRL in this context and also delves into the challenges associated with it. Additionally, it discusses prospective future directions for the application of DRL in the scheduling of cloud computing resources. This examination offers a detailed insight into the evolving role of DRL in optimizing and streamlining scheduling processes within cloud environments.

In study [65], researchers propose a novel technique named RATS-HM (Resource Allocation and Task Scheduling using Hybrid Machine Learning) for combined resource allocation and efficient task scheduling in cloud computing, aimed at overcoming existing challenges. This technique includes several key components: First, it features an improved cat swarm optimization algorithm-based short scheduler for task scheduling, known as ICS-TS, which is designed to minimize make-span time while maximizing throughput. Secondly, it incorporates a group optimization-based deep neural network (GO-DNN) that facilitates efficient resource allocation, taking into account various design constraints such as bandwidth and resource load. Thirdly, the technique introduces a lightweight authentication scheme called NSUPREME for data encryption, enhancing the security of data storage. Finally, to establish the effectiveness of the RATS-HM technique, simulations are conducted in different setups, and the results are compared with state-of-the-art techniques. Notably, these simulations demonstrate that the proposed technique significantly reduces power consumption, achieving a reduction of nearly 68%, thereby highlighting its potential in enhancing energy efficiency in cloud computing environments.

Workload scheduling is another critical factor in VM management, as it affects the performance and energy consumption of VMs. Several algorithms have been proposed to optimize workload scheduling in VMs. In [37], the authors proposed an algorithm that schedules VMs based on their resource usage and energy consumption. The algorithm uses a genetic algorithm to find the optimal schedule of VMs, which reduces energy consumption and improves performance.

Several algorithms have been proposed to optimize workload scheduling in containers. In [56], the authors proposed an algorithm that schedules containers based on their resource usage and energy consumption. The algorithm uses a genetic algorithm to find the optimal schedule of containers, which reduces energy consumption and improves performance.

4.3. CONSOLIDATION

Consolidation involves moving VMs and containers across servers to consolidate workloads and turn off idle servers. This approach is based on the observation that many servers in data centers are often idle, leading to wasted energy consumption. By consolidating workloads and turning off idle servers, dynamic workload consolidation can reduce energy consumption in VM and container management.

Consolidation algorithms are pivotal in optimizing resource utilization and reducing energy consumption within cloud computing environments. These algorithms effectively tackle the issue of underutilized resources. They achieve this by dynamically migrating and packing Virtual Machines (VMs) onto a smaller number of physical machines. This strategy leads to a decrease in the total count of active machines, subsequently lowering their collective energy footprint. This approach not only enhances efficiency but also contributes significantly to the sustainability of cloud computing operations.

Table 3 shows focus, algorithms and key features of consolidation studies between 2018-2023. The study [70] proposes a DVS-based consolidation algorithm that balances energy savings with performance requirements. Another study [40], introduces a two-phase algorithm considering both energy consumption and server temperature to prevent overheating. In the study [55] the ant colony metaheuristic is utilized to schedule VMs while minimizing energy consumption and migration overhead. The study [52] ensures timely execution of critical tasks by incorporating deadline guarantees into the consolidation process. Another study [51] addresses fairness concerns by allocating resources among tenants proportionally to their usage and priorities. The study [49] considers both execution time and resource costs when scheduling containerized workloads for optimal performance and budget efficiency. Another study [48] employs deep learning to predict resource demands and dynamically consolidate VMs for improved performance and resource utilization. The study [65] which is applied for situational awareness and monitoring is also used to combine reinforcement learning and heuristics to optimize scheduling and resource allocation in cloud environments. In another study, referenced as [66], collaborative learning is enabled among scheduling agents within a dynamic cloud environment, facilitating efficient resource management.

Paper	Focus	Approach	Key Features
[70]	Energy-aware	DVS-based	Balances energy savings and performance with dynamic voltage scaling
[40]	Energy-aware & Temperature-aware	Two-phase	Minimizes energy consumption and prevents overheating
[42]	Energy-aware & Performance aware	Evolutionary game theory	Optimization of energy consumption
[32]	Energy-aware	Energy efficient strategy	Highest possible throughput and energy consumption
[33]	Energy-aware	Coalitional-game	Partitioning and Shutting down PMs
[38]	Energy-aware	Reinforcement learning	Optimizes the power management forVMs
[55]	Energy-aware	Ant colony optimization	Schedules VMs while minimizing energy and migration overhead
[52]	Performance-aware & QoS-aware	Deadline-aware	Guarantees timely execution of critical tasks with deadlines
[51]	Performance-aware & Fairness-aware	Priority-aware	Allocates resources among tenants proportionally to their usage and priorities

[49]	Performance-aware & Cost-aware	Cost-aware container scheduling	Optimizes execution time and resource costs for containerized workloads
[48]	Machine Learning	Deep reinforcement learning	Predicts resource demands and dynamically consolidates VMs for improved performance and utilization
[65]	Machine Learning (Hybrid)	Hybrid reinforcement learning	Combines reinforcement learning and heuristics for optimal scheduling and resource allocation
[66]	Machine Learning (Federated)	Federated learning	Enables collaborative learning among scheduling agents for dynamic cloud resource management

Table 3. Consolidation studies between 2018-2023 in respect of energy consumption

In study [42], a new approach is proposed to address the consolidation problem of Virtual Machines (VMs) with the aim of optimizing their energy consumption. This approach includes the introduction of a novel algorithm, complemented by an energy consumption model specifically designed to aid the algorithm in identifying optimal solutions. Experimental results from this study indicate that significant reductions in energy consumption can be achieved through dynamic VM consolidation. When compared with other classic algorithms, the proposed algorithm demonstrates superior performance, achieving an average energy consumption savings of 42%. This highlights its effectiveness in enhancing energy efficiency in VM management.

The research presented in [32] introduces an Energy-Efficient Strategy (EES) designed for consolidating Virtual Machines (VMs) in a cloud environment. The primary objective of this strategy is to achieve a reduction in energy consumption while concurrently handling an increased volume of tasks with the highest possible throughput. Central to this proposal is the utilization of the performance-to-power ratio, which is employed to establish upper thresholds for overload detection. Additionally, EES takes into account the overall data center workload utilization to determine lower thresholds, an approach aimed at minimizing the frequency of virtual machine migrations. Simulation results from the study indicate that EES effectively leads to energy-efficient workload consolidation. This is achieved with a minimal number of migrations and a reduction in overall energy consumption. The results further suggest that EES is capable of saving energy without compromising the requirements of the user's workload. Specifically, the strategy is shown to reduce energy consumption by 19%, demonstrating its effectiveness in enhancing energy efficiency in cloud computing environments.

In the approach proposed in [33], the methodology begins by partitioning Physical Machines (PMs) into distinct groups according to their load levels. Following this, a coalitional-game-based VM consolidation algorithm, referred to as CGMS, is employed. This algorithm is instrumental in selecting members from these groups to form efficient coalitions. Subsequently, VM migrations are executed among the members of these coalitions to maximize each coalition's payoff. Additionally, the approach involves shutting down

PMs that exhibit low energy efficiency, further optimizing the overall system performance. Experimental results based on multiple cases clearly demonstrate that the proposed approach achieves higher energy-saving (32.30% higher than Sercon in three scenarios on average).

Hypervisors like VMware ESXi and Microsoft Hyper-V have built-in power management features that can reduce energy consumption. These features include dynamic voltage and frequency scaling, which adjusts processor power consumption based on workload demands.

A method about energy-efficient VM management is power management, which involves dynamically adjusting the power consumption of servers and VMs based on workload demand. Several algorithms have been proposed to optimize power management in VMs. In the study referenced as [24], the authors explore the advantages of virtualization in the context of power management within data centers. They propose a power-aware Virtual Machine (VM) consolidation algorithm. This algorithm is designed to consolidate VMs onto a fewer number of physical servers, aiming to reduce energy consumption while simultaneously adhering to performance requirements. In [38], the authors proposed an algorithm that optimizes the power management of VMs using a reinforcement learning approach. The algorithm learns the optimal power management strategy for VMs, which reduces energy consumption and improves performance.

4.4. ALLOCATION

Resource allocation algorithms play a crucial role in cloud computing, determining how resources like CPU, memory, and storage are assigned to competing tasks and applications. As cloud environments evolve, researchers are constantly developing new and optimized allocation algorithms to address specific challenges and objectives. Researchers are concentrating their studies on specific areas of allocation, which include focusing on energy consumption, resource distribution, and budgetary limitations.

Table 4 shows focus, algorithms and key features of allocation studies between 2018-2023. The study [58] proposes an ant colony optimization-based approach for heterogeneous cloud environments, prioritizing energy savings. Another study [61] introduces a priority queue-based algorithm that guarantees a minimum QoS for tenants in multi-tenant cloud environments. The study [60] integrates deadline constraints into the allocation process, prioritizing timely execution of critical tasks. The study [62] considers both execution time and resource costs when scheduling containerized workloads, aiming for cost-efficient performance. The study [34] addresses deadline and budget constraints simultaneously for workflow tasks using reinforcement learning. Another hybrid approach [65] combines reinforcement learning and heuristics for joint task scheduling and resource allocation. Another study [66] proposes a federated learning framework for dynamic cloud environments, enabling agents to learn from each other and adjust allocation strategies.

Paper	Focus	Approach	Key Features
[54]	Energy-aware & Performance improvement	Genetic algorithm	Find the optimal allocation to reduce energy consumption
[19]	Request priority & Allocation rate	Priority-aware resource allocation	Prioritize VM requests and parition the hosts
[30]	Dynamic virtual machine allocation & Energy-aware	Power-aware scheduling-based resource allocation	Maintain the total CPU utilization

[58]	Energy-aware & Heterogeneous	Ant colony optimization	Prioritizes energy savings while considering resource heterogeneity
[61]	Fairness-aware & Multi-tenant	Priority queues	Guarantees minimum QoS for tenants through queue management
[60]	Fairness-aware & Deadline-constrained	Deadline-aware	Prioritizes timely execution of critical tasks
[62]	Cost-aware & Containerized workloads	Cost-aware container scheduling	Optimizes execution time and resource costs for containers
[34]	Cost-aware & Deadline & Budget-constrained	Reinforcement learning	Considers deadlines, budget, and workflow tasks for optimal allocation
[65]	Hybrid (Reinforcement learning & heuristics)	Joint task scheduling and resource allocation	Combines techniques for optimal allocation in both areas
[66]	Federated learning	Dynamic environments & Collaborative learning	Enables agents to learn from each other and adapt allocation strategies

Table 4. Allocation studies between 2018-2023 in respect of energy consumption

In [54], the authors proposed an algorithm that optimizes the allocation of CPU and memory resources in containers. The algorithm uses a genetic algorithm to find the optimal allocation of resources, which reduces energy consumption and improves performance. In [34], the authors proposed an algorithm that dynamically adjusts the CPU frequency of containers based on their workload. The algorithm monitors the CPU utilization of containers and adjusts the CPU frequency accordingly, which reduces energy consumption without affecting the performance of containers.

A recent solution in the field of cloud computing involves allocating high-priority Virtual Machine (VM) requests in close proximity, while other requests are assigned using the First Fit Decreasing (FFD) method. However, this approach does not distinguish between different types of requests by partitioning the hosts. Addressing this gap, study [19] introduces the Priority-Aware Resource Allocation (PARA) algorithm. PARA categorizes VM requests into three types: highly critical, critical, and normal. Correspondingly, it partitions the hosts into three levels, allowing high-priority VM requests access to a relatively larger pool of hosts compared to lower-priority requests. To evaluate the effectiveness of the PARA algorithm, simulations were conducted with 10,000 to 100,000 VM requests and 20 to 30 hosts. These were then compared with a non-partitioned pool of hosts using the FFD method, focusing on allocation rate and weighted score in relation to serving requests. The study specifically calculated the total allocated differences between PARA and FFD for both 20 and 30 hosts. The results show that PARA offers a more favorable total allocated difference, where the ratio is y:x:1, with $y > 1$ and $x \geq 1$, indicating its superior performance in effectively allocating resources according to the priority of requests.

In study [30], researchers developed a genetic heuristic search optimization technique for dynamic consolidation of Virtual Machines (VMs), which is based on adaptive utilization thresholds. This approach aims to ensure a high level of compliance with Service Level Agreements (SLA). The strategy involves setting upper and lower utilization thresholds for hosts, with the primary goal being to maintain the total CPU utilization by all VMs within these dynamically fluctuating thresholds.

To address the challenges of the dynamic virtual machine allocation policy, the power-aware scheduling-based resource allocation method, termed G-PARS, was proposed. This method was designed to optimize the allocation of VMs, focusing on energy efficiency and quality of service (QoS).

Experimental results from the study indicate that the G-PARS strategy outperforms particle swarm optimization strategies in several aspects. It not only maintains high QoS but also results in lower energy consumption. Additionally, G-PARS demonstrates a significant advantage in reducing the number of active hosts, particularly under workloads that are life-threatening or critically demanding. In terms of power reduction, G-PARS shows an improved performance in reducing the energy consumption of each data center, with reductions ranging between 13-22% compared to existing algorithms. This highlights the effectiveness of G-PARS in achieving both energy efficiency and high-quality service in cloud computing environments.

4.5. PLACEMENT/MIGRATION

Placement and migration algorithms play a crucial role in optimizing resource utilization and performance in cloud computing. They determine where virtual machines (VMs) should be placed and when they should be migrated between physical machines to ensure efficient resource allocation, meet service level agreements (SLAs), and minimize operational costs.

Table 5 shows focus, algorithms and key features of placement/migration studies between 2018-2023. A study [48] leverages deep reinforcement learning to predict resource demands and dynamically place VMs for improved performance and resource utilization. It demonstrates the effectiveness of DRL in proactive cloud resource management. Another study [55] proposes an ant colony optimization-based approach for heterogeneous cloud environments, prioritizing energy savings while considering resource heterogeneity. A study [40] introduces a two-phase algorithm that prioritizes energy savings while preventing overheating in cloud environments, addressing thermal concerns during VM migration. An approach proposed in [59] is VMP-A3C, which is a dynamic placement algorithm based on A3C via migration. the purpose of placement is to find the best mapping between VMs and HMs, considering some constraints and objectives. In [36], the authors proposed an algorithm that optimizes the placement of VMs on physical servers to reduce energy consumption. The algorithm uses a heuristic approach to find the optimal placement of VMs, which reduces energy consumption without affecting the performance of VMs.

In study [39], researchers introduced a novel approach to VM consolidation, which takes into account both the current and anticipated future utilization of resources. This approach is operationalized through two key components: the host overload detection (UP-POD) and host underload detection (UP-PUD). A Gray-Markov-based model is employed to accurately predict future resource utilization.

The effectiveness of this proposed approach was tested using real-world workload traces in CloudSim, and the results were benchmarked against existing algorithms. The simulations demonstrated that the approach significantly reduces both the number of VM migrations and energy consumption, while still maintaining the Quality of Service (QoS) guarantee.

In terms of energy consumption, the proposed approach achieved an average reduction of 42.7%, 38.1%, 39%, and 33.1% compared to the THR (Thresholding), IQR (Interquartile Range), MAD (Median Absolute Deviation), and LR (Local Regression) approaches, respectively. This substantial improvement is attributed to the reductions in both energy consumption and the Service Level Agreement Violation (SLAV) rate. The results are particularly noteworthy as they indicate that the approach effectively balances the trade-off between power cost and the assurance of QoS. This balance is crucial in cloud computing environments, where both energy efficiency and service reliability are key priorities.

Also, power management techniques can be applied to container management. In the study [27], the authors proposed a power management framework for containerized applications that uses dynamic voltage and frequency scaling (DVFS) to adjust the power consumption of CPUs based on workload demand. The framework also includes a container migration algorithm to balance the energy consumption of different physical servers.

Paper	Focus	Approach	Key Features
[48]	VM placement	Deep reinforcement learning	Minimize energy consumption, improve performance
[50]	VM placement/ migration	Dynamic best-fit decreasing	Resource utilization, Energy saving
[55]	VM placement	Ant colony optimization	Prioritize energy savings, consider resource heterogeneity
[40]	VM migration	Two-phase algorithm	Minimize energy consumption, prevent overheating
[59]	VM placement	Dynamic placement algorithm based on A3C	Control and monitor the migration process.
[36]	VM placement	Heuristic approach	Optimal placement of VM with energy consumption without affecting performance of VMs
[57]	Container placement	Heuristic approach	Energy consumption without affecting the performance of containers
[39]	VM Migration & Energy-aware	Host overload detection, Host underload detection, Gray-Markov	Reducing VM migation and energy consumption
[27]	Container migration	Dynamic voltage and frequency scaling	Power management framework for containerized apps.
[31]	Microservice placement	Fine-tuned Sunflower Whale Optimization Algorithm	Resource utilization
[43]	Energy-aware	Minimization of Migration	Dynamic thresholding mechanism, replacing static thresholds

[45]	Energy-aware	Cooperative Coevolution Genetic Programming	Dynamic Resource Allocation and Consolidation
[46]	Energy-aware & VM placement	Multi-objective evolutionary algorithm	Reduce resource wastage, power consumption and network transmission delay

Table 5. Placement/Migration studies between 2018-2023 in respect of energy consumption

To address the critical challenge of resource wastage, power consumption, and network transmission delay in cloud computing, an efficient optimization methodology is essential for cloud providers. The study by [46] proposes a novel approach utilizing NSGA-III, a multi-objective evolutionary algorithm, to simultaneously minimize these three crucial objectives and achieve a non-dominated solution. The efficacy of the proposed NSGA-III algorithm is rigorously evaluated by comparing its performance metrics, namely Overall Nondominated Vector Generation (ONVG) and Spacing, against established multi-objective algorithms like VEGA, MOGA, SPEA, and NSGA-II. The outcome reveals a significant performance advantage for the proposed algorithm, demonstrating a 7% improvement in ONVG and a 12% improvement in Spacing compared to the best existing method. Moreover, the robustness of these findings is further verified through statistical validation using ANOVA and DMRT tests.

Focusing on the multifaceted challenges of microservices in cloud computing, particularly meeting end-user demands while adhering to stringent Service Level Agreements (SLAs), [31] proposes a novel QoS-aware resource allocation model based on the Fine-tuned Sunflower Whale Optimization Algorithm (FSWOA). This model optimizes microservice deployment on physical machines, maximizing resource utilization and achieving improved efficiency. Through comprehensive experimental simulations, the authors demonstrate the proposed approach's superiority over established baseline methods, including SFWOA, Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO). Notably, their model achieves statistically significant reductions in key performance metrics, including execution time (up to 4.26%), memory consumption (up to 11.29%), CPU utilization (up to 17.07%), and service cost (up to 24.22%).

In the study detailed in [45], the researchers propose a novel approach: a hybrid Cooperative Coevolution Genetic Programming (CCGP) hyper-heuristic. This method is designed to autonomously generate heuristics that are effective for addressing the dynamic Resource Allocation and Consolidation (RAC) problem. Diverging from existing methodologies, this approach combines the Best Fit algorithm with automatically designed heuristics to effectively address the two interdependent sub-problems of RAC. Additionally, the researchers introduce a new energy model that provides a more accurate depiction of energy consumption. This model is more aligned with realistic settings, accounting for factors like real-world workload patterns and the heterogeneity of Physical Machines (PMs). Experimental results from this study indicate that this approach can significantly reduce energy consumption when compared to two state-of-the-art methods.

In study [43], a prediction mechanism has been adopted and implemented in conjunction with the existing Minimization of Migration (MM) policy, specifically tailored for large history data sets. This is complemented by a dynamic thresholding mechanism, replacing static thresholds. Rigorous simulations were conducted to test this approach, and the results demonstrate a reduction in energy consumption in cloud data centers.

In [57], the authors proposed an algorithm that optimizes the placement of containers on physical hosts to reduce energy consumption. The algorithm uses a heuristic approach to find the optimal placement of containers, which reduces energy consumption without affecting the performance of containers.

5. CONCLUSION

Figure 4 shows that the maximum energy consuming equipment in cloud computing are servers. Virtualization allows multiple virtual machines (VMs) to share the same physical resources, which leads to better utilization of hardware and reduces the overall energy consumption. Containerization also facilitates improvements in various aspects of technology management. It enhances application portability, accelerates the cycles of application development and deployment, and leads to better resource utilization and scalability. This is because containers can be managed and orchestrated more efficiently compared to traditional virtual machines.

In Section IV, we discussed various algorithms proposed in recent research for optimizing energy consumption in VM and container management. These algorithms include dynamic workload consolidation, power management, resource allocation, and container orchestration. These algorithms can efficiently allocate resources, manage workloads, and optimize cooling and power distribution systems to reduce energy consumption in VM and container management. The selection of a suitable algorithm depends on the specific requirements of the system and the workload characteristics.

Furthermore, the effectiveness of these algorithms can be further enhanced by combining them with other optimization techniques, such as workload prediction, resource provisioning, and power management.

There is a novel approach within the domain of virtualization management that proposes the integration of multiple stages. For instance, some studies have introduced algorithms that simultaneously encompass situational awareness and allocation, representing a consolidated strategy under the umbrella of the virtualization manager.

Container orchestration involves managing the deployment, scaling, and monitoring of containerized applications. This approach is based on the observation that containerization can reduce energy consumption by reducing the overhead of virtualization. By efficiently managing containerized applications, container orchestration algorithms can further reduce energy consumption in VM and container management.

In light of these findings, our survey-based research further investigates methods for optimizing CPU utilization, with a specific focus on the roles of virtualization and containerization in cloud computing. The survey encompassed a diverse range of studies and expert opinions, revealing an emergent trend in the increased research focus on containerization technologies. This shift indicates a growing interest in exploring containerization as an alternative or complementary approach to traditional virtualization, especially in the context of its potential to improve energy efficiency in server operations. Containerization's increasing prominence and its potential to offer enhanced efficiency and scalability, which are crucial for reducing energy consumption in cloud computing servers.

In summary, containerization is emerging as a key technology in cloud computing, offering enhanced efficiency and scalability. These features are instrumental in reducing energy consumption in cloud computing servers, thereby addressing one of the major challenges in the field. The shift towards containerization indicates a significant step towards more sustainable and energy-efficient cloud computing practices.

This comprehensive survey thus provides valuable insights into current trends and future directions in CPU utilization strategies, particularly in the rapidly evolving field of cloud computing.

REFERENCES

1. Jones, N. How to stop data centres from gobbling up the world’s electricity. Nature 2018, 561, 163. (Reference)
2. Andrae, A.S.G.; Edler, T. On Global Electricity Usage of Communication Technology: Trends to 2030. Challenges 2015, 6, 117-157.
3. CBRE. Global Data Center Trends 2023. (Reference)
4. Singh, S., Bhardwaj, S., Pandey, H., Beniwal, G. (2021). Anomaly Detection Using Federated Learning. In: Bansal, P., Tushir, M., Balas, V., Srivastava, R. (eds) Proceedings of International Conference on Artificial Intelligence and Applications. Advances in Intelligent Systems and Computing, vol 1164. Springer, Singapore. https://doi.org/10.1007/978-981-15-4992-2_14
5. Trends in data center energy consumption under the European Code of Conduct for data center energy efficiency, Castellazzi L, Avgerinou M, Bertoldi P, European Commission. 2017,
6. Kenga, D., Omwenga, V.O., & Ogao, P.J., “Virtual Machine Customization Using Resource Using Prediction for Efficient Utilization of Resources in IaaS Public Clouds”, Journal of Information Technology and Computer Science, 2021, https://api.semanticscholar.org/CorpusID:239461887}.
7. P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the Art of Virtualization. SIGOPS Oper. Syst. Rev. 37, 164–177. 2003.
8. D. Bernstein. Containers and cloud: From LXC to Docker to kubernetes. IEEE Cloud Computing 1, 3 (2014), 81–84. 2014.
9. Dec 2019. KVM [online]. (Dec 2019). (Reference)
10. C. D. Graziano. A performance analysis of Xen and KVM hypervisors for hosting the Xen Worlds Project. 2011.
11. Dec 2019.Containers at Google [online]. (Reference)
12. Dec 2019. LXC [online]. (Reference)
13. Dec 2019. Docker [online] (Reference)
14. P.H. Kamp and R. NM. Watson. Jails: Confining the omnipotent root. In The 2nd International SANE Conference, Vol. 43. 116. 2000.
15. Dec 2019. Window container [online]. (Reference)
16. M. G. Xavier, M. V. Neves and C. A. F. D. Rose. 2014. A Performance Comparison of Container-Based Virtualization Systems for MapReduce Clusters. 2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing. Torino, 2014, pp. 299-306. doi: 10.1109/PDP.2014.78 .
17. Renugadevi, T., K. Geetha, Natarajan Prabaharan, and Pierluigi Siano. 2020. "Carbon-Efficient Virtual Machine Placement Based on Dynamic Voltage Frequency Scaling in Geo-Distributed Cloud Data Centers" Applied Sciences 10, no. 8: 2701. https://doi.org/10.3390/app10082701 .
18. J. Kon, N. Mizusawa, A. Umezawa, S.Yamaguchi, and J.Tao. 2017. Highly Consolidated Servers with Container-based Virtualization. IEEE International Conference on Big Data. 2017.
19. Kushagra Kinger, Ajeet Singh, and Sanjaya Kumar Panda. 2022. Priority-Aware Resource Allocation Algorithm for Cloud Computing. In Proceedings of the 2022 Fourteenth International Conference on

Contemporary Computing (IC3-2022). Association for Computing Machinery, New York, NY, USA, 168–174. https://doi.org/10.1145/3549206.3549236

20. Jon Watson. 2008. VirtualBox: bits and bytes masquerading as machines. Linux J. 2008, 166, pages
21. F. Paraiso, S. Challita, Y. Al-Dhuraibi and P. Merle. 2016. Model- Driven Management of Docker Containers. 2016 IEEE 9th International Conference on Cloud Computing (CLOUD). San Francisco, CA, 2016, pp. 718-725. doi: 10.1109/CLOUD.2016.0100
22. L.Youseff, R.Wolski, B.Gorda, and C.Krintz. 2006. Paravirtualization for HPC systems. In Proceedings of the 2006 international conference on Frontiers of High Performance Computing and Networking (ISPA'06), Geyong Min, Beniamino Martino, Laurence T. Yang, Minyi Guo, and Gudula Rünger (Eds.). Springer-Verlag, Berlin, Heidelberg, 474-486.
23. W. Jiye, Z. Biyu, Z. Fa, S. Xiang, Z. Nan, and L. Zhiyong, ‘‘Data center energy consumption models and energy efficient algorithms,’’ J. Comput. Res. Develop., vol. 56, no. 8, p. 1587, 2019.
24. I. Foster, R. Montero, B. Sotomayor and I. Llorente, "Virtual Infrastructure Management in Private and Hybrid Clouds" in IEEE Internet Computing, vol. 13, no. 05, pp. 14-22, 2009. doi: 10.1109/MIC.2009.119.
25. M. Moravcik, P. Segec, M. Kontsek, J. Uramova and J. Papan, "Comparison of LXC and Docker Technologies," 2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA), Košice, Slovenia, 2020, pp. 481-486, doi: 10.1109/ICETA51985.2020.9379212.
26. J. -B. Wang et al., "A Machine Learning Framework for Resource Allocation Assisted by Cloud Computing," in IEEE Network, vol. 32, no. 2, pp. 144-151, March-April 2018, doi: 10.1109/MNET.2018.1700293
27. Feng Li, Wen Jun Tan, Wentong Cai, A wholistic optimization of containerized workflow scheduling and deployment in the cloud–edge environment, Simulation Modelling Practice and Theory, Volume 118, 2022, 102521, ISSN 1569-190X, https://doi.org/10.1016/j.simpat.2022.102521.
28. Mohammad S. Islam, William Pourmajidi, Lei Zhang, John Steinbacher, Tony Erwin, and Andriy Miranskyy. 2021. Anomaly detection in a large-scale cloud platform. In Proceedings of the 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP '21). IEEE Press, 150–159. https://doi.org/10.1109/ICSE-SEIP52600.2021.00024
29. N. Moretti, M. Rutren and T. Bessell, "Space Situational Awareness Sensor Tasking: A Comparison Between Step-Scan Tasking and Dynamic, Real-Time Tasking," 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 2018, pp. 1339-1344, doi: 10.23919/ICIF.2018.8455851.
30. Journal, IJCSMC. “An Heuristics Based Dynamic Power-Aware Resource Allocation for Cloud Computing.” IJCSMC 7, no. 11 (2018): 204–15.
31. Kumar, M, Samriya, JK, Dubey, K, Gill, SS. QoS-aware resource scheduling using whale optimization algorithm for microservice applications. Softw: Pract Exper. 2023; 1-20. doi: 10.1002/spe.3211.
32. Saadi, Y., El Kafhali, S. Energy-efficient strategy for virtual machine consolidation in cloud environment. Soft Comput 24, 14845–14859 (2020). https://doi.org/10.1007/s00500-020-04839-2.
33. Xiao, X. et al. (2019). A Novel Coalitional Game-Theoretic Approach for Energy-Aware Dynamic VM Consolidation in Heterogeneous Cloud Datacenters. In: Miller, J., Stroulia, E., Lee, K., Zhang, LJ. (eds) Web Services – ICWS 2019. ICWS 2019. Lecture Notes in Computer Science(), vol 11512. Springer, Cham. https://doi.org/10.1007/978-3-030-23499-7_7.

34. Mboula, J.E., Kamla, V.C., Hilman, M.H., & Djamégni, C.T., "Energy-efficient workflow scheduling based on workflow structures under deadline and budget constraints in the cloud", 2022, https://arxiv.org/abs/2201.05429 .
35. H. A. Salam, F. Davoli, A. Carrega and A. Timm-Giel, "Towards Prediction of Power Consumption of Virtual Machines for Varying Loads," 2018 28th International Telecommunication Networks and Applications Conference (ITNAC), Sydney, NSW, Australia, 2018, pp. 1-6, doi: 10.1109/ATNAC.2018.8615319.
36. Peng, Zhihao & Barzegar, Behnam & Yarahmadi, Maryam & Motameni, Homayun & Pirozmand, Poria. (2020). Energy-Aware Scheduling of Workflow Using a Heuristic Method on Green Cloud. Scientific Programming. 2020. 14. 10.1155/2020/8898059.
37. Rostami, S., Broumandnia, A. & Khademzadeh, A. An energy-efficient task scheduling method for heterogeneous cloud computing systems using capuchin search and inverted ant colony optimization algorithm. J Supercomput (2023). https://doi.org/10.1007/s11227-023-05725-y .
38. K. Dubey, S. C. Sharma and A. A. Nasr, "A Simulated Annealing based Energy-Efficient VM Placement Policy in Cloud Computing," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.119.
39. Sun-Yuan Hsieh, Cheng-Sheng Liu, Rajkumar Buyya, Albert Y. Zomaya, "Utilization-prediction-aware virtual machine consolidation approach for energy-efficient cloud data centers", Journal of Parallel and Distributed Computing, 139,2020, 99-109,ISSN 0743-7315, https://doi.org/10.1016/j.jpdc.2019.12.014,
40. Yavari, M., Ghaffarpour Rahbar, A. & Fathi, M. Temperature and energy-aware consolidation algorithms in cloud computing. J Cloud Comp 8, 13 (2019). https://doi.org/10.1186/s13677-019-0136-9
41. Guangyao Z,Wenhong T and Rajkumar B, "Deep Reinforcement Learning-based Methods for Resource Scheduling in Cloud Computing: A Review and Future Directions", 2021, https://arxiv.org/abs/2105.04086.
42. Liu, X, Wu, J, Chen, L, Zhang, L. Energy-aware virtual machine consolidation based on evolutionary game theory. Concurrency Computat Pract Exper. 2022; 34(10):e6830. doi:10.1002/cpe.6830.
43. Bhattacherjee, S., Das, R., Khatua, S. et al. Energy-efficient migration techniques for cloud environment: a step toward green computing. J Supercomput 76, 5192–5220 (2020). https://doi.org/10.1007/s11227-019-02801-0.
44. Tahseen Khan, Wenhong Tian, Guangyao Zhou, Shashikant Ilager, Mingming Gong, and Rajkumar Buyya. 2022. Machine learning (ML)-centric resource management in cloud computing: A review and future directions. J. Netw. Comput. Appl. 204, C (Aug 2022). https://doi.org/10.1016/j.jnca.2022.103405.
45. Chen Wang, Hui Ma, Gang Chen, Victoria Huang, Yongbo Yu, and Kameron Christopher. 2023. Energy-Aware Dynamic Resource Allocation in Container-Based Clouds via Cooperative Coevolution Genetic Programming. In Applications of Evolutionary Computation: 26th European Conference, EvoApplications 2023, Held as Part of EvoStar 2023, Brno, Czech Republic, April 12–14, 2023, Proceedings. Springer-Verlag, Berlin, Heidelberg, 539–555. https://doi.org/10.1007/978-3-031-30229-9_35.

46. Gopu, A., Thirugnanasambandam, K., R, R. et al. Energy-efficient virtual machine placement in distributed cloud using NSGA-III algorithm. J Cloud Comp 12, 124 (2023). https://doi.org/10.1186/s13677-023-00501-y.
47. Hooman Alavizadeh, Julian Jang-Jaccard, Simon Yusuf Enoch, Harith Al-Sahaf, Ian Welch, Seyit A. Camtepe, and Dan Dongseong Kim. 2022. A Survey on Cyber Situation-awareness Systems: Framework, Techniques, and Insights. ACM Comput. Surv. 55, 5, Article 107 (May 2023), 37 pages. https://doi.org/10.1145/3530809
48. Stefanini, Matteo & Lancellotti, Riccardo & Baraldi, Lorenzo & Calderara, Simone., 2019. A Deep Learning based approach to VM behavior identification in cloud systems, https://doi.org/10.48550/arXiv.1903.01930
49. Zhu, L., Huang, K., Fu, K. et al. A priority-aware scheduling framework for heterogeneous workloads in container-based cloud. Appl Intell 53, 15222–15245 (2023). https://doi.org/10.1007/s10489-022-04164-1
50. Neha Gupta, Kamali Gupta, Deepali Gupta, Sapna Juneja, Hamza Turabieh, Gaurav Dhiman, Sandeep Kautish, Wattana Viriyasitavat, "Enhanced Virtualization-Based Dynamic Bin-Packing Optimized Energy Management Solution for Heterogeneous Clouds", Mathematical Problems in Engineering, vol. 2022, Article ID 8734198, 11 pages, 2022. https://doi.org/10.1155/2022/8734198
51. B. Liang, X. Dong and X. Zhang, "A Power-aware Scheduling Algorithm in Multi-tenant IaaS Clouds," 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 2019, pp. 250-253, doi: 10.1109/SIPROCESS.2019.8868410
52. S. Luo, P. Fan, H. Xing and H. Yu, "Meeting Coflow Deadlines in Data Center Networks With Policy-Based Selective Completion," in IEEE/ACM Transactions on Networking, vol. 31, no. 1, pp. 178-191, Feb. 2023, doi: 10.1109/TNET.2022.3187821
53. Y. Li et al., "Virtual Machine Performance Analysis and Prediction," 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), Sharjah, United Arab Emirates, 2020, pp. 1-5, doi: 10.1109/CCCI49893.2020.9256518.
54. Sida Xing, Feihu Han, Suiyang Khoo, “Extreme-Long-short Term Memory for Time-series Prediction”, 2022, https://doi.org/10.48550/arXiv.2210.08244
55. W. Wei, H. Gu, W. Lu, T. Zhou and X. Liu, "Energy Efficient Virtual Machine Placement With an Improved Ant Colony Optimization Over Data Center Networks," in IEEE Access, vol. 7, pp. 60617-60625, 2019, doi: 10.1109/ACCESS.2019.2911914
56. Radhakrishnan Balu, Ajinkya Borle, “Bayesian Networks based Hybrid Quantum-Classical Machine Learning Approach to Elucidate Gene Regulatory Pathways”, 2019, https://doi.org/10.48550/arXiv.1901.10557
57. Alfred Ultsch, Jörg Hoffmann, Maximilian Röhnert, Malte Von Bonin, Uta Oelschlägel, Cornelia Brendel, Michael C. Thrun, “An Explainable AI System for the Diagnosis of High Dimensional Biomedical Data” 2021, https://doi.org/10.48550/arXiv.2107.01820
58. Harvinder S., Sanjay T., Pardeep K., Sukhpal S. G., Rajkumar B., "Metaheuristics for scheduling of heterogeneous tasks in cloud computing environments: Analysis, performance evaluation, and future directions", Simulation Modelling Practice and Theory, Volume 111, 2021, 102353, ISSN 1569-190X, https://doi.org/10.1016/j.simpat.2021.102353.
59. P. Wei, Y. Zeng, B. Yan, J. Zhou, E. Nikougoftar, VMP-A3C: Virtual machines placement in cloud computing based on asynchronous advantage actor-critic algorithm, Journal of King Saud University

- Computer and Information Sciences, Volume 35, Issue 5, 2023, 101549, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2023.04.002

60. F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, N. T. Hieu and H. Tenhunen, "Energy-Aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model," in IEEE Transactions on Cloud Computing, vol. 7, no. 2, pp. 524-536, 1 April-June 2019, doi: 10.1109/TCC.2016.2617374.
61. Fahad M, Shojafar M, Abbas M, Ahmed I, Ijaz H. A multi-queue priority-based task scheduling algorithm in fog computing environment. Concurrency Computat Pract Exper. 2022; 34(28):e7376. doi:10.1002/cpe.7376
62. Rodriguez, M. & Buyya, R., Container Orchestration With Cost-Efficient Autoscaling in Cloud Computing Environments. In B. Gupta & D. Gupta (Eds.), Handbook of Research on Multimedia Cyber Security, 2020, pp. 190-213, IGI Global, https://doi.org/10.4018/978-1-7998-2701-6.ch010
63. Qiu, Yifei & Wu, Shaohua & Wang, Ying. , "On the Scheduling Policy for Multi-process WNCS under Edge Computing", 2021, https://doi.org/10.48550/arXiv.2109.12535.
64. Rostami, S., Broumandnia, A. & Khademzadeh, A. An energy-efficient task scheduling method for heterogeneous cloud computing systems using capuchin search and inverted ant colony optimization algorithm. J Supercomput (2023). https://doi.org/10.1007/s11227-023-05725-y
65. Bal, P.K.; Mohapatra, S.K.; Das, T.K.; Srinivasan, K.; Hu, Y.-C. A Joint Resource Allocation, Security with Efficient Task Scheduling in Cloud Computing Using Hybrid Machine Learning Techniques. Sensors 2022, 22, 1242. https://doi.org/10.3390/s22031242
66. Y. Sun, S. Zhou, Z. Niu and D. Gündüz, "Dynamic Scheduling for Over-the-Air Federated Edge Learning With Energy Constraints," in IEEE Journal on Selected Areas in Communications, vol. 40, no. 1, pp. 227-242, Jan. 2022, doi: 10.1109/JSAC.2021.3126078.
67. Roland C. F., Marc I., Anthony N. C., Martin J. ., "Scalable Circuits for Preparing Ground States on Digital Quantum Computers: The Schwinger Model Vacuum on 100 Qubits", 2023, https://doi.org/10.48550/arXiv.2308.04481
68. Banerjee, Suman and Tekawade, Atherve, 2023, "Cost and Reliability Aware Scheduling of Workflows Across Multiple Clouds with Security Constraints". http://dx.doi.org/10.2139/ssrn.4469649
69. Zhou, G., Tian, W., & Buyya, R. (2021). Deep Reinforcement Learning-based Methods for Resource Scheduling in Cloud Computing: A Review and Future Directions. ArXiv, abs/2105.04086.
70. Zhou, Qiheng & Xu, Minxian & Gill, Sukhpal Singh & Gao, Chengxi & Tian, Wenhong & Xu, Chengzhong & Buyya, Rajkumar. (2020). Energy Efficient Algorithms based on VM Consolidation for Cloud Computing: Comparisons and Evaluations. 489-498. 10.1109/CCGrid49817.2020.00-