Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Distributed Resource Management in Cloud Computing: A Review of Allocation, Scheduling, and Provisioning Techniques

Nabeel N. Ali[1], Subhi R. M. Zeebaree[2]
nabeel.ali@dpu.edu.krd[1], subhi.rafeeq@dpu.edu.krd[2]
[1] IT Dept., Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq
[2] Energy Eng. Dept., Technical College of Engineering, Duhok Polytechnic University, Duhok, Iraq

Article Information

Submitted : 9 Mar 2024
Reviewed: 14 Mar 2024
Accepted : 1 Apr 2024

Keywords

cloud computing, resource management, allocation, scheduling, provisioning

Abstract

This review paper provides an in-depth examination of distributed resource management in cloud computing, focusing on the critical elements of allocation, scheduling, and provisioning. Cloud computing, characterized by its dynamic and scalable nature, necessitates efficient resource management techniques to optimize performance, cost, and service. The study encompasses a comprehensive analysis of various strategies in resource allocation, scheduling methodologies, and provisioning techniques within the cloud computing paradigm. Through comparative analysis, this paper aims to highlight the synergies and trade-offs inherent in these methods, offering a holistic view of distributed resource management. It contributes to the field by bridging the gap in existing literature, presenting a critical, comparative analysis of current strategies and their interplay in distributed cloud environments.

A. Introduction

In the rapidly evolving landscape of cloud computing, distributed resource management stands as a cornerstone, pivotal to the efficient and effective operation of cloud-based services. This paper delves into the intricate world of resource management within the cloud, focusing on the critical aspects of allocation, scheduling, and provisioning. These elements are fundamental in orchestrating the complex interplay of resources that are often dispersed across multiple data centers globally [1].

The emergence of cloud computing has revolutionized the way we perceive and interact with digital infrastructure. At its core, it involves the delivery of various computing services — including servers, storage, databases, networking, software, and more — over the Internet. Such services are hosted in expansive data centers, where a plethora of IT equipment operates in unison to ensure seamless service delivery[2], [3]. These centers, acting as the backbone of cloud computing, are intricate arrays of routers, switches, servers, and other critical devices, all requiring meticulous management to maintain operational efficiency [4].

A critical challenge in this domain is the management of distributed resources, a task that becomes increasingly complex due to the sheer scale and dynamic nature of cloud environments. The efficient allocation of these resources, the strategic scheduling of tasks, and the proactive provisioning of infrastructure are fundamental to optimizing performance and cost-effectiveness in cloud computing. The heterogeneity of resources, coupled with fluctuating and unpredictable workloads, necessitates sophisticated management strategies to harness the full potential of cloud computing [5].

Despite the abundance of research in this field, there remains a significant gap in comprehensive, critical analyses of these management techniques. Existing literature often tackles these topics in isolation or without an in-depth comparative analysis of the various approaches. This paper seeks to fill this gap by providing a detailed review of the current strategies in resource allocation, scheduling, and provisioning within the cloud computing paradigm. We aim to not only present a comparative analysis of these techniques but also to discuss their synergies and trade-offs in the context of distributed cloud environments [6].

This review is structured to first establish a foundational understanding of the key concepts in cloud computing and distributed resource management. It then progresses to an in-depth analysis of resource allocation strategies, followed by a critical examination of scheduling methodologies and provisioning techniques. Through this comprehensive approach, the paper aims to offer valuable insights and a holistic view of distributed resource management in cloud computing, addressing both its current state and future possibilities [7].

The remainder of the paper is structured as follows: Section 2 delves into the background and fundamental concepts of cloud computing, setting the stage for a comprehensive understanding of distributed resource management. In Section 3, we thoroughly examine various resource allocation strategies, exploring their nuances and implications within the cloud computing sphere. Section 4 is dedicated to an in-depth analysis of scheduling techniques, highlighting their role and impact in cloud computing. The focus of Section 5 is on resource provisioning

and management, where we discuss the methods and challenges involved. Section 6 addresses the contemporary challenges and opportunities in cloud resource management, identifying emerging trends and potential research directions. In Section 7, we present a comparative analysis and discussion of the various methodologies and strategies explored in the previous sections. The paper concludes with Section 8, where we summarize our findings and offer insights into future directions for research in distributed resource management in cloud computing.

B. Background Theory

a. Cloud Computing Basics

The core infrastructure of cloud computing predominantly resides in geographically dispersed data centers, accessible to users via the network. Each data center encompasses a physical space housing an array of IT equipment arranged in racks. These include disk enclosures, servers, routers, switches, load balancers, firewalls, and various other devices crucial for data processing and storage [8].

Central to the data center's functionality is the effective monitoring and management of this equipment. Monitoring systems play a pivotal role, tracking usage and issuing alerts for abnormal activity, ensuring stability and efficiency in operations. For instance, as illustrated in Figure 1, the network topology of a data center is a critical component that underpins its operational integrity.

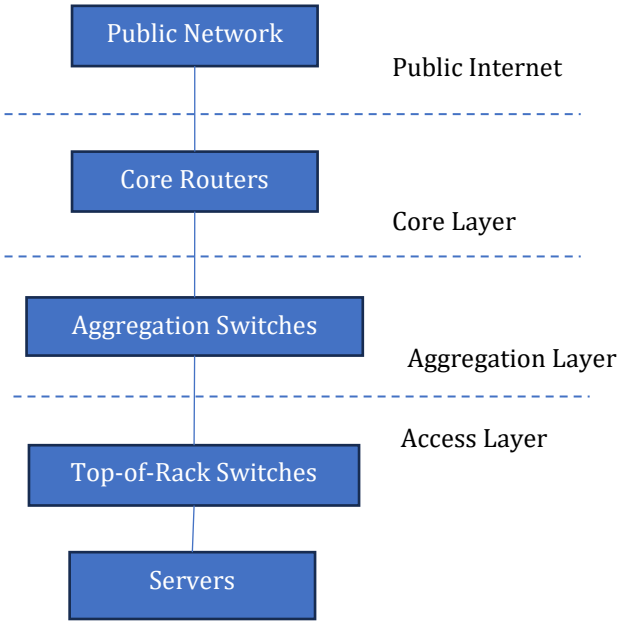

Figure 1. Representation of data center network topology.

The architectural design of racks, with servers stacked vertically, optimizes space utilization and simplifies cabling processes. The Top of Rack (ToR) design connects to Aggregation Switches (AS), ensuring redundancy and robust connectivity for each server unit. The aggregation layer plays a key role, channeling traffic from multiple ToR switches to the core layer, where core routers

handle high-speed data transfer, maintaining the backbone of internet connectivity [9].

Furthermore, the deployment of load balancers is instrumental in enhancing network bandwidth. By distributing connection requests across multiple servers, they optimize data traffic management. Underpinning all these elements is the server functionality, with different servers catering to specific needs like file sharing, email services, and web hosting [10].

This intricate infrastructure reflects core cloud computing principles such as scalability, elasticity, and on-demand resource availability, highlighting the dynamic and flexible nature of cloud services.

b. Distributed Systems

At the heart of cloud computing lies the intricate architecture of distributed systems (see Figure 2), networks of autonomous computers functioning cohesively as a unified entity [11]. These systems are the cornerstone of the cloud's ability to scale efficiently and maintain high availability[12]. Horizontal scalability, a hallmark feature, allows for the integration of additional nodes to handle increased demand, embodying the cloud's flexible, pay-as-you-go model [13], [14]. Additionally, the distribution of resources across various nodes and geographic locations enhances fault tolerance, ensuring service continuity even in the event of individual node failures. This is particularly critical for applications where reliability is paramount.

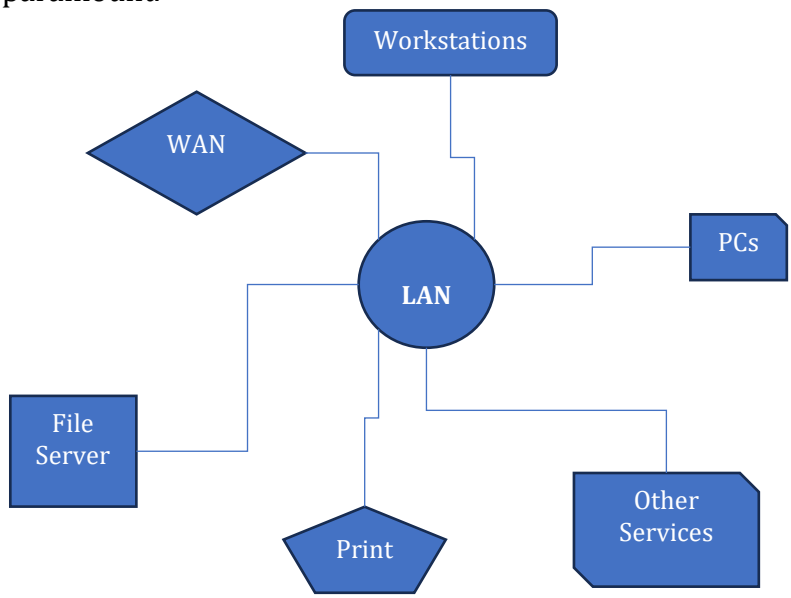

Figure 2. Distributed System

The efficiency and effectiveness of cloud services are further influenced by the choice of distributed computing models. The client-server model, forming the backbone of Software as a Service (SaaS) and Platform as a Service (PaaS), typifies the conventional cloud architecture with centralized servers serving multiple clients. In contrast, Peer-to-Peer (P2P) networks, where each node functions both as a client and a server, exemplify a more egalitarian approach, commonly employed in distributed storage systems. Grid computing, though often associated with cloud computing, differentiates itself by focusing on fewer, more intensive

computational tasks. These models collectively contribute to the cloud's versatility, catering to a diverse range of computational needs and applications [15].

Navigating the challenges inherent in distributed systems is pivotal for the advancement of cloud computing [16]. Key issues include maintaining efficient communication between nodes, particularly over large distances, which involves addressing latency and bandwidth limitations. Additionally, ensuring robust security protocols is crucial to safeguard data privacy and integrity as information traverses' public networks[17]. A significant technical hurdle is also presented by the need for real-time data synchronization across various nodes, ensuring consistency and reliability in rapidly evolving cloud environments [18].

C. Literature Review

In the intricate ecosystem of cloud computing, effective resource management is pivotal, encompassing the integral processes of resource allocation, scheduling, and provisioning as illustrated in Figure 3. Resource allocation is at the forefront, involving the strategic distribution of computational resources like processing power and storage to various applications and services, aiming to optimize infrastructure usage and balance demand. Alongside, scheduling plays a critical role in maintaining operational efficiency; it orchestrates the sequence and timing of tasks, considering priorities and resource needs, thereby maximizing throughput and enhancing user experience. Complementing these is the dynamic process of provisioning, which is the adaptive allocation and reallocation of resources in response to fluctuating demands. This adaptive mechanism is essential for the cloud's hallmark features of scalability and elasticity, enabling responsive scaling of resources to ensure cost-effective and efficient utilization. Collectively, these processes form the backbone of cloud computing, ensuring that it remains robust, agile, and capable of meeting diverse and evolving user needs.

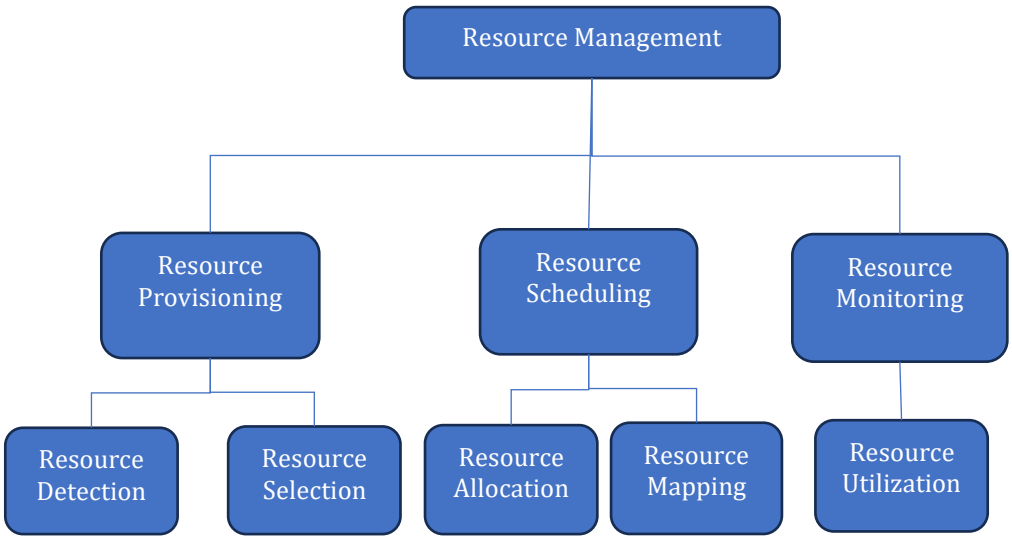

Figure 3. Taxonomy of Resource-Management in Cloud

a. Categorization of Cloud Computing Resources

Cloud computing resources are broadly categorized into three types: physical, virtual, and logical resources. Physical resources encompass components

like processors, memory, disk drives, Network Interface Controllers (NICs), peripheral devices (such as keyboards), networking products, storage media, and other tangible elements. These resources are strategically distributed in data centers worldwide to efficiently serve cloud clients [19].

Virtualization, a key technique in cloud computing, partitions computer resources into multiple execution environments, enhancing the productivity of physical machines. This allows for running multiple instances of operating systems and applications on a single physical machine. Virtual resources, which include virtual CPUs (vCPUs), virtual memory, virtual switches (vSwitches), and virtual storage (vSAN), are dynamically shared, scheduled, and scaled based on consumer demand [20].

Resource Management Systems (RMS) play a crucial role in managing these resources. They monitor resource availability, ensuring they meet user requests, and continuously optimize resource allocation. Data centers host a variety of applications, ranging from web servers and databases to customer business apps. With the shift towards Infrastructure as a Service (IaaS), virtual resource management tools have become essential for provisioning resources on-demand, irrespective of the location of available computing resources [21].

Logical resource management involves system abstractions that temporarily control physical resources. This includes managing aspects like operating systems, energy [22], network throughput/bandwidth, information security, protocols, APIs, network loads, and delays [23]. These abstractions ensure effective control and optimization of the underlying physical resources.

b. The Requirements of Resource Management in Cloud Computing

Cloud computing is characterized by five distinct features, each with its specific requirements for resource management:

Table 1: Characteristics of Resource Management and Their Requirements in Cloud Computing

Characteristics	Requirements	Objectives
On-demand self-service		
Broad network access	Intelligent and business-related resource management	
Resource pooling	End-to-end resource chain management	
Rapid elasticity	Dynamic deployment management of virtual resources	QoS & Cost optimization
Measured service	Dynamic adaptive resource management	
On-demand self-service	Monitoring and reporting of resource usage at an appropriate level	

c. Challenges in Resource Management within Cloud Computing

Managing resources in cloud computing presents several challenges. This section briefly reviews key issues in cloud resource management. When cloud users request resources, the provider's Resource Management System (RMS) schedules and allocates these resources, but this process faces hurdles like minimizing energy consumption and costs while maximizing performance.

Efficient strategies identified in the literature [24], [25] include energy efficiency, bandwidth cost reduction, and optimizing storage and performance. This review primarily focuses on challenges related to power, networking, compute, and storage resources.

Power usage is a significant concern in cloud resource management [26]. Idle servers in data centers consume about 60% of power, with the remainder used by cooling and security systems. Strategies like turning off idle servers and automating cooling systems are advised for improved energy efficiency. Other solutions include Dynamic Component Deactivation (DCD), Dynamic Performance Scaling (DPS), and Dynamic Voltage and Frequency Scaling (DVFS) [27], with DVFS being particularly effective in maximizing energy savings and user profits. An energy-aware algorithm supporting DVFS is discussed in [28], and a multi-agent-based resource management approach for minimizing energy consumption is proposed in [29].

Network resources also pose challenges. Studies [30], [31] suggest VM resource scheduling algorithms that focus on minimizing network bandwidth costs. The optimization of computing resources is crucial, as highlighted by [32], who presented a dynamic programming approach for this purpose. Traditional storage resource management systems encounter numerous obstacles in cloud environments, indicating a need for modern systems adept at managing cloud storage resources efficiently.

d. Resource Management Techniques

Resource management in cloud computing primarily aims to minimize service costs and enhance performance, security, and energy efficiency. Various techniques are evaluated based on several parameters, including monetary aspects (like service cost), application performance metrics (such as response time, execution time, delay, SLA violations, task type, required processor number, throughput, resource availability, and utilization), security, and energy efficiency (overall power and energy consumption) [33]. Optimization methods like load balancing, Round Robin, Bin Packing algorithm, and Gradient Search algorithm are recognized for improving performance, reducing costs, and lowering energy consumption in IaaS resources [34].

Researches [35], [36] have explored a stochastic model focusing on load balancing and scheduling within cloud computing clusters. Another approach presented in [37] involves using Self-Organizing Clouds (SOC) to maximize resource utilization and achieve optimal execution times. Table 5 provides a comparative summary of cloud resources based on selected metrics.

i. Resource Allocation

Resource allocation in cloud computing is the process of assigning appropriate resources to consumer tasks for efficient completion. This typically involves allocating a virtual machine that meets the consumer's specified requirements. Users submit tasks with specific time constraints, and a key aspect is determining how these tasks are allocated to virtual machines. Effective resource allocation depends on several factors: the resources allocated, the time required, the sequence of actions, and their interdependencies [38]. Additionally, it

encompasses the discovery, selection, provisioning, scheduling, and management of resources, entailing decisions on the timing, type, location, and quantity of resources allocated to consumers, as illustrated in Figure 4.

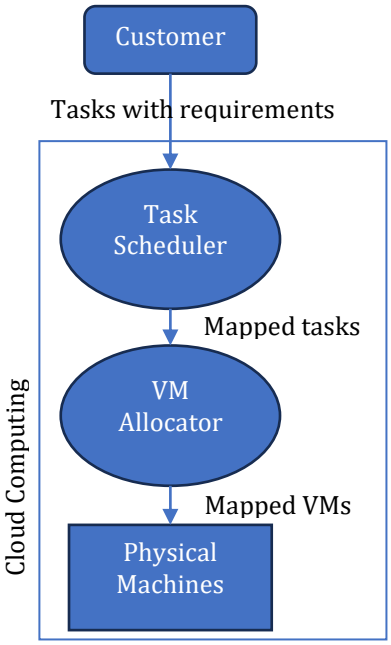

Figure 4. Resource Allocation Layers in Cloud Computing.

Figure 5 outlines the general steps in resource allocation: consumers submit requests to the resource allocator, requests are queued, the allocator informs the allocation unit, which then requests resources from the Infrastructure as a Service (IaaS). If available, IaaS responds positively, a Virtual Machine (VM) is created from the VM pool, and the allocator is informed. The requests are then dequeued, and resources are allocated.

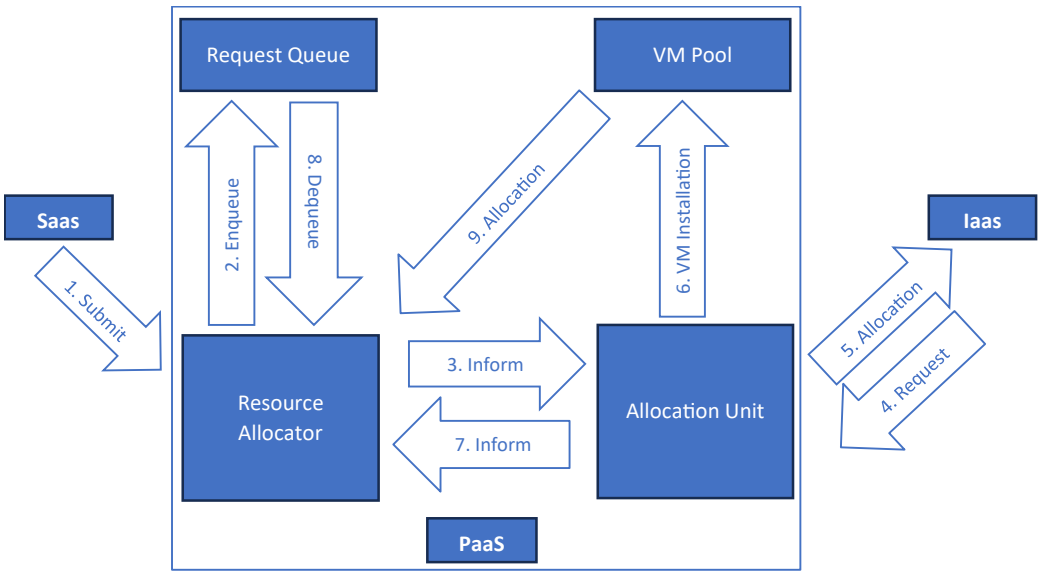

Figure 5. The basic flow of resource allocation in cloud computing.

A key challenge in this process is the lack of information sharing between cloud service providers and consumers. Providers typically do not disclose details about their resources, and consumers do not reveal specifics about their application workloads. This lack of transparency hinders the optimization of resource allocation [39]. The unpredictability of user requests, running on data centers over the internet, further complicates resource allocation in cloud computing. Challenges include predicting consumer and application demands, ensuring physical machines can meet the resource needs of all VMs, providing efficient networking services with quality QoS, and developing auction-based allocation mechanisms that benefit both providers and consumers. Additionally, minimizing SLA violations while maximizing resource utilization is crucial, as QoS often impacts resource allocation strategies to reduce costs [40].

Resource allocation in cloud computing employs various methods to efficiently utilize resources and meet consumer needs. These techniques are classified into several categories: strategic (adapting to changing consumer demands), target resources (focusing on specific requested resources), auction (bidding for resources), optimization (enhancing resource efficiency), scheduling (prioritizing tasks), and power (allocating resources with minimal power consumption), as depicted in Figure 6. Each category is evaluated based on critical parameters from both cloud service provider and consumer perspectives.

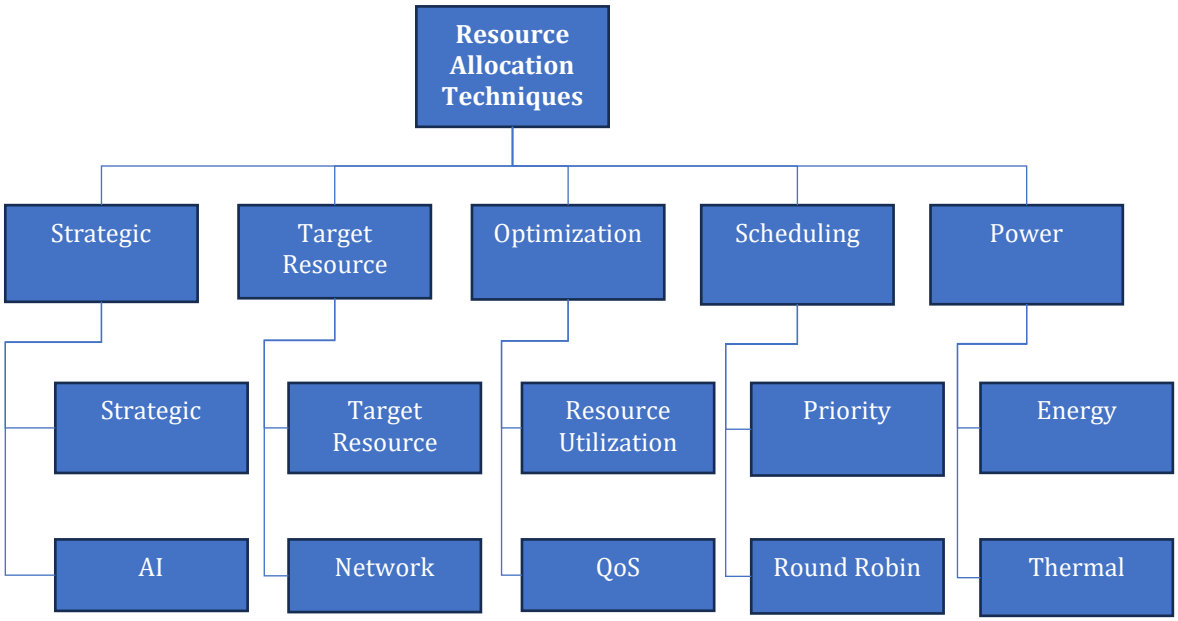

Figure 6. Taxonomy of resource allocation in cloud computing.

Key parameters include cost, which is crucial for providers in determining the financial efficiency of services; resource utilization, focusing on optimal use of resources to prevent idleness and reduce data center costs; power, addressing the growing need for energy-efficient services; and workload, reflecting the system's capacity to process tasks efficiently. For both providers and consumers, execution time (minimizing task completion time), response time (speed of system responses), and user satisfaction (meeting consumer expectations) are vital.

Other important factors include the Quality of Service (QoS) and Service Level Agreement (SLA) adherence for both parties, fraud prevention, and revenue generation. The effectiveness of these parameters is often rated on a scale from 1 to 5, with 5 being the highest. However, a high value is not always ideal; for instance, lower values are preferable for cost, response time, execution time, workload, and power. Conversely, high values are desirable for user satisfaction, SLA fulfillment, resource utilization, fraud prevention, and revenue.

The assessment of each feature in resource allocation techniques involves a consistent protocol: defining the feature, discussing relevant literature, and evaluating the feature based on the mentioned parameters.

Ref.	Objectives	Major Concepts	Limitations	Future Work
[41]	Review resource allocation techniques in cloud computing, presenting a taxonomy and analyzing articles in each category	Resource allocation strategies and challenges in cloud computing	Not explicitly stated, but generally include the scope of literature review and methodological constraints.	Focus on artificial intelligence for scheduling, green optimization of data centers, and exploring mobility patterns for task assignment in cloud computing
[42]	To address limitations in mobile learning applications using cloud computing, focusing on solving scalability issues.	Cloud scalability, integration of mobile learning with cloud computing, and resource allocation.	Scalability in cloud computing for mobile learning applications.	explore parallel reinforcement learning techniques in other contexts like smart grids.
[43]	Focus on optimizing a multi-tier computation system considering allowable latency and bandwidth constraints	Employ a practical wireless Heterogeneous Networks (HetNets) approach for scalable computation offloading	Increased energy consumption as a potential drawback of offloading	Ensure availability of adequate radio resources and refine offloading selection for better energy efficiency
[44]	To jointly optimize offloading strategy and resource allocation in Cloud Radio Access Network (C-RAN) with Mobile Edge Computing (MEC) for profit maximization.	Task-aware C-RAN with MEC, spectrum efficiency-based offloading, Lagrangian multiplier method for resource allocation	Scalability and effectiveness in diverse network scenarios might be limited.	Further exploration of efficient resource allocation methods and extending the model to different network architectures
[45]	Decrease the waiting time for tasks and enhance overall resource utilization	Strategy involving the assignment of tasks with the shortest implementation time to the quickest completing resources	Current resource management limited to a single cloud environment	Propose expansion of virtual machines to multiple cloud environments for broader resource distribution
[46]	Effectively leverage on-demand resources with implications for fuel economy, vehicle	Address resource allocation challenges in automotive systems	Distinct management models for public (decentralized) and private (centralized,	Develop comprehensive resource provisioning models applicable to both public and private clouds

	comfort, and safety in automotive systems	under both public and private cloud models	auction-based) clouds	
[47]	aims to review innovative resource allocation methods in computing environments, focusing on methods based on artificial intelligence	Includes a comprehensive literature study on machine learning and deep learning methods for resource allocation in computing environments.	restricted application of certain techniques in real-world systems, particularly in relation to parameters like reaction time and handling time	suggests exploring deep reinforcement learning and convolutional neural network algorithms combined with meta-heuristic algorithms for resource allocation in cloud computing environments

Table 2: Comparison of Resource Allocation Approaches in Cloud Computing

ii. Scheduling Strategies

The increasing popularity of cloud computing has sparked significant interest among researchers in the field of resource scheduling theory. Resource scheduling essentially involves mapping each user request to a suitable resource, taking into account quality standards and ensuring efficient utilization of all available resources without breaching Service Level Agreements (SLAs). The cloud scheduling process generally encompasses three stages:

1) Resource Finding and Sorting – The datacenter service broker identifies available resources from the pool and assesses their status.
2) Resource Selection – Resources are chosen based on required parameters, finalizing decisions for resource allocation.
3) Request Submission – The selected resources are deployed to execute the user's requested task (refer to Figure 7).

Various methods focus on optimizing multiple quality attributes in resource scheduling [48], including cost, makespan, execution cost and time, response time, bandwidth/speed, priority, workload, availability, throughput, reliability, recovery time, SLA, and utilization. These methods include batch mode, load-based, auction-based, agent-based, credit-based, and dynamic resource scheduling [49], each tailored to different types of resource requests in cloud computing.

However, research in resource scheduling techniques remains limited. Studies have explored challenges and opportunities in cloud computing resource scheduling [50], with approaches such as a batch mode scheduling scheme using the Berger model for effective resource allocation [7], and frameworks considering penalty costs for SLA violations to maximize provider profit and resource utilization [51]. Additionally, load-based resource scheduling methods have been proposed to enhance system performance, particularly under heavy traffic conditions [52].

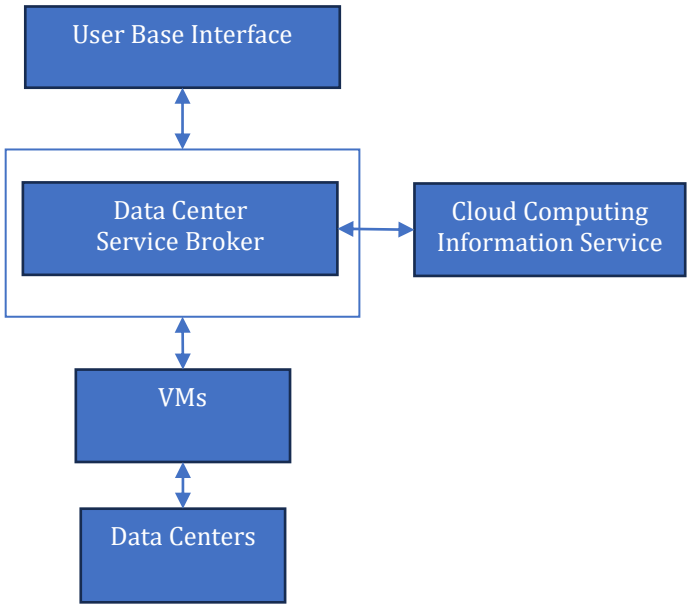

Figure 7. Resource scheduling in cloud computing.

A variety of resource scheduling techniques have been formulated, each focusing on specific scheduling attributes. These attributes include Fault Tolerance, Execution Time, Response Time, Load Balancing, Makespan, Throughput, Resource Utilization, Scalability, Quality of Service, and Performance [53]. This section delves into several resource scheduling techniques, examining them through the lens of these distinct scheduling attributes.

In cloud computing, resource scheduling is a crucial area of focus, with numerous techniques being developed based on various scheduling attributes. The Batch Mode Dynamic Scheduling Algorithm, for instance, alternates between online and batch modes, catering to different request rates, with batch mode scheduling requests only after a thorough analysis of collected sets [34]. Similarly, the Load-Based Scheduling Algorithm is designed to maximize profit and efficiency under strict deadline constraints, specifically addressing heavy-tailed requests through effective workload control [54].

Another significant approach is the Active Resource Provisioning Algorithm, which emphasizes balanced load distribution and server sharing, utilizing a method called 'Skewness' to evaluate resources in VMs [55]. The Aggressive Resource Provisioning Algorithm, or SPRNT, maintains high-level Quality of Service (QoS) by dynamically adjusting VM instances, proving particularly effective in handling rapidly increasing workloads [56].

In high-demand scenarios, the Auction-Based Resource Allocation Algorithm is used to minimize resource wastage, using auctions for dynamic resource allocation, thus optimizing revenue [34]. The Autoscaling Prediction Model for Resource Provisioning employs a predictive framework to anticipate workload and provision VMs accordingly, using various predictive techniques such as ARIMA, Neural Networks (NN), and Support Vector Machine (SVM) [57].

The Elasticity-Based Scheduling Heuristic Algorithm (EBSH) compares favorably against algorithms like ACO and HBO, especially in terms of execution time due to VM reusability [58]. The Enriched Workflow Scheduling Algorithm

(EWSA) optimizes execution time and cost for dependent tasks, providing more realistic scheduling for QoS attributes compared to CSO techniques [59].

Each of these techniques offers unique advantages and caters to different operational requirements in cloud computing resource scheduling (see Table 3).

Table 3: Comparison of Resource Scheduling Algorithms Based on Scheduling Attributes

Scheduling Method	Year	Throughput Capable	Supports Makespan	Utilizes Resources	Balances Load	Enhances Performance	Scalable	Quality Service Oriented	Cost-Effective
Dynamic Scheduling (Batch Mode) [60]	2022	Yes	Yes	Yes	No	No	No	No	No
Algorithm Based on Load [52]	2017	No	No	No	Yes	Yes	No	No	Yes
Resource Allocation (Auction Method) [61]	2016	No	No	No	No	No	No	No	Yes
Scheduling Model (Autoscaling Prediction) [62]	2023	No	No	No	Yes	Yes	Yes	Yes	No
Heuristic (Elasticity-Based Scheduling) [58]	2023	No	No	No	No	No	No	No	Yes
Workflow Scheduling (Enhanced) [59]	2022	No	No	No	Yes	No	No	No	No
Scheduling System (Intelligent Agent-Based) [57]	2023	Yes	Yes	Yes	Yes	No	No	No	No
Allocation Scheme (Dynamic Resources) [34]	2022	No	No	Yes	Yes	Yes	Yes	No	No
Algorithm (Credit-Based Scheduling) [63]	2022	No	No	No	Yes	No	No	No	Yes

iii. Provisioning Methods

A major challenge in cloud computing resource provisioning is balancing user costs with service provider resource utilization [64]. Cloud computing has become a popular choice for enterprises due to its ability to effectively manage software (such as database servers, load balancers) and hardware resources (like CPU, storage, and network). This management ensures optimal application performance, taking into account Service Level Agreements (SLAs) which guarantee Quality of Service (QoS) parameters including performance, availability, reliability, and response time.

Effective resource provisioning, whether static or dynamic, alongside appropriate allocation, is essential to utilize resources efficiently without breaching SLAs and satisfying QoS requirements. Both over-provisioning and under-provisioning of resources should be avoided. Power consumption is another critical factor, necessitating strategies to minimize power use and optimize VM placement, thereby avoiding excessive energy use.

The ultimate goal for cloud users is cost minimization through resource rental, while service providers aim to maximize profits through efficient resource

allocation. To meet these objectives, users must request resources from providers, specifying whether they need static or dynamic provisioning. This ensures that providers understand the required resource instances and types for specific applications. Successful provisioning should achieve key QoS parameters like availability, throughput, security, response time, reliability, and performance, all within the confines of agreed-upon SLAs. The common resource provisioning techniques are:

Static Provisioning: This technique is suitable for applications with predictable and stable demands. In static provisioning, customers arrange with providers for services in advance, leading to the preparation of necessary resources before the service starts. Billing is typically handled through a flat fee or monthly charges.

Dynamic Provisioning: Ideal for applications with fluctuating demands, dynamic provisioning involves the migration of Virtual Machines (VMs) to different compute nodes within the cloud as needed. This method allows providers to allocate additional resources when required and retract them when they are not in use. Billing is based on a pay-per-use model. When applied to create a hybrid cloud, this approach is often referred to as cloud bursting.

User Self-provisioning: Also known as cloud self-service, this approach enables customers to acquire resources directly from the cloud provider via a web interface. Customers create an account, purchase resources, and make payments using a credit card. The resources provided by the cloud service become accessible to the customer within a very short timeframe, often within minutes or hours s depicted in Figure 8.

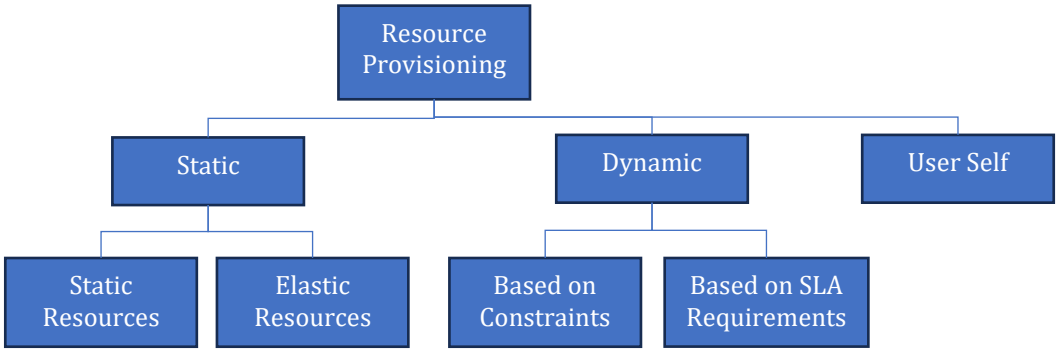

Figure 8. Resource Provisioning Strategies

Table 4: Comparative Analysis of Cloud Computing Resource Provisioning Techniques

Ref.	Year	Technique	Advantages	Limitations
[65]	2022	Deadline-Oriented Resource Allocation in Hybrid Clouds.	Efficient allocation of diverse resources, leading to reduced execution times for applications.	Inapplicable for high-performance, data-intensive applications.
[66]	2012	Dynamic Resource Allocation in Multi-Tenant Service Clouds.	Aligns services offered by tenants with the specific needs of clients.	Ineffective for practical testing in real-world, multi-domain cloud systems.
[67]	2022	Elastic Application Container for Cloud Resource	Excels in providing resource efficiency and adaptability.	Limited to Java programming, not suitable for web-based

		Optimization.		applications.
[68]	2022	Hybrid Cloud Resource Management in the Event of Resource Failures.	Adapts to user workload models, allowing flexible strategy selection based on quality of service, performance needs, and budget constraints.	Infeasible for conducting actual experiments.
[69]	2022	Efficient VM Request Handling with Placement Constraints in IaaS Clouds.	Enhances runtime efficiency and revenue maximization through effective VM-to-PM mappings.	Impractical for managing medium to large-scale issues.
[70]	2020	Failure-Resilient Resource Management in Hybrid Cloud Infrastructures.	Enhances quality of service by significantly reducing deadline violation rates and slowdowns at minimal costs on public clouds.	Unable to conduct real experiments; struggles with VM transfers between public and private clouds for local infrastructure failure management.
[71]	2021	VM Provisioning for Enhanced Profit and Reduced SLA Violations.	Minimizes SLA violations and boosts profitability.	Aggravates resource allocation and load balancing challenges across datacenters.
[72]	2021	Risk-Aware VM Consolidation and Resource Aggregation.	Drastically reduces the need for server resources, allowing for deactivation of excess servers.	Only considers CPU needs for VM allocation.
[73]	2021	Semantic Resource Management in Inter-Cloud Environments.	Maximizes customer satisfaction by augmenting resources in a federated cloud, addressing interoperability issues.	Needs to meet specific QoS parameters such as response time and throughput for interactive applications.
[74]	2023	Adaptive Power-Aware VM Provisioning Using Swarm Intelligence.	Optimizes VM placement, leading to substantial power savings.	Not ideal for power conservation in contemporary data centers.
[75]	2022	Optimal Resource Allocation for Cloud Environments.	Effectively allocates resources for SaaS users within budget and deadline constraints, optimizing QoS.	Limited to SaaS users and providers only.
[76]	2023	Adaptive Provisioning for Read-Intensive Multi-Tier Cloud Applications.	Automates identification and resolution of bottlenecks in multi-tiered web applications.	Not suitable for clustered n-tier applications in cloud environments.

D. Discussion and Comparison

In the analysis of the provided table comparing different resource management techniques over a range of years, distinct trends and patterns emerge across the various metrics: Quality of Service (QoS), Utilization Rate, Cost, Energy Efficiency, Scalability, and Security.

Resource consolidation techniques, as referenced in [77] and [78], consistently show positive outcomes in Cost, Energy Efficiency, Scalability, and Security, but they do not improve QoS and Utilization Rate. This suggests these techniques prioritize operational efficiency and sustainability over performance.

Resource adaptation [79] and Resource provisioning [80] display a mixed performance. While excelling in Energy Efficiency and Scalability, they have little impact on QoS and Utilization Rate, indicating a selective focus in their approach that may prioritize long-term benefits over immediate performance enhancements.

Different patterns are observed in Resource allocation [81] and Resource discovery [82], where each excels in either Utilization Rate or Scalability but not uniformly across other metrics. This could reflect a more targeted approach, focusing on specific aspects of resource management.

Recent developments such as Cloud Resource Elasticity [83], Advanced Resource Optimization [21], and AI-Driven Resource Management [84] show comprehensive improvement across all metrics. These techniques, incorporating the latest technologies, mark a significant evolution in resource management, achieving an equilibrium between cost, efficiency, scalability, and security without compromising service quality or utilization.

The table reflects the ongoing evolution in resource management techniques. Earlier methods often focused on specific areas like cost-efficiency, sometimes at the expense of service quality or security. In contrast, recent advancements, especially those employing AI and cloud technologies, demonstrate a holistic approach, addressing a broad spectrum of metrics effectively. This evolution signifies a shift towards more balanced and comprehensive resource management strategies.

Table 5: Comparative Analysis on Different Resource Management Techniques

Ref.	Techniques	QoS	Utilization Rate	Cost	Energy Efficiency	Scalability	Security
[77]	Resource consolidation	No	No	Yes	Yes	Yes	Yes
[78]		No	No	Yes	Yes	Yes	Yes
[85]		No	No	No	No	No	No
[79]	Resource adaptation	No	No	No	Yes	Yes	Yes
[86]		No	No	Yes	Yes	Yes	No
[87]		Yes	Yes	Yes	Yes	No	Yes
[81]	Resource allocation	No	Yes	No	Yes	Yes	No
[88]		Yes	Yes	Yes	No	Yes	No
[89]		No	No	Yes	No	No	No
[80]	Resource provisioning	No	No	No	Yes	No	Yes
[90]		No	Yes	No	No	Yes	Yes
[91]		No	No	Yes	Yes	No	No
[82]	Resource discovery	No	Yes	No	No	No	No
[92]		Yes	No	No	No	Yes	No
[93]		No	Yes	No	No	No	No
[94]	Resource scheduling	Yes	Yes	Yes	Yes	Yes	Yes
[95]		No	Yes	Yes	Yes	Yes	Yes
[96]		Yes	Yes	Yes	No	No	Yes
[97]	Resource blocking	No	Yes	No	Yes	Yes	No
[98]		No	No	No	No	No	No
[38]	Modeling and Resource estimation	No	No	Yes	Yes	No	Yes
[99]		No	No	No	No	Yes	Yes
[100]		No	No	No	No	No	No
[101]	Resource mapping	Yes	No	No	Yes	Yes	No
[102]		No	No	Yes	Yes	No	Yes

[103]		No	No	No	No	No	No
[83]	Cloud Resource Elasticity	Yes	Yes	Yes	Yes	Yes	Yes
[21]	Advanced Resource Optimization	Yes	Yes	Yes	Yes	Yes	Yes
[84]	AI-Driven Resource Management	Yes	Yes	Yes	Yes	Yes	Yes

E. Extracted Statistics

The evaluation of various resource management techniques in the domain of cloud computing discloses distinct trends and patterns which are pivotal in understanding the evolution of these methodologies. These patterns emerge across diverse metrics such as Quality of Service (QoS), Utilization Rate, Cost, Energy Efficiency, Scalability, and Security.

1. **Resource Consolidation Techniques:** Investigations into these techniques, as noted in sources [51] and [52], reveal consistent positive impacts in terms of Cost, Energy Efficiency, Scalability, and Security. However, these techniques do not demonstrate improvements in QoS and Utilization Rate. This finding suggests a tendency for these methods to prioritize operational efficiency and sustainability over immediate performance metrics.
2. **Resource Adaptation and Provisioning**: As per sources [53] and [54], these methods show a mixed range of performances. They excel in Energy Efficiency and Scalability but demonstrate minimal influence on QoS and Utilization Rate. This indicates a selective focus in their approach, potentially prioritizing long-term benefits over immediate performance enhancements.
3. **Resource Allocation and Discovery:** Referring to sources [55] and [56], distinct patterns are observed where each technique excels in either Utilization Rate or Scalability, but not uniformly across other metrics. This could be indicative of a more targeted approach, focusing on specific elements of resource management.
4. **Recent Developments:** Advanced techniques such as Cloud Resource Elasticity [57], Advanced Resource Optimization [1], and AI-Driven Resource Management [58] exhibit comprehensive improvement across all evaluated metrics. These methodologies, by incorporating cutting-edge technologies, signify a substantial evolution in resource management, achieving a balance between cost efficiency, scalability, and security without compromising on service quality or utilization rates.

These findings reflect the ongoing progression in resource management techniques, transitioning from earlier methods that often concentrated on specific areas like cost-efficiency, sometimes at the expense of service quality or security, to recent advancements that exhibit a more holistic approach, effectively addressing a broad spectrum of metrics.

F. Recommendations

The implications of this comprehensive review are significant, offering insights for both practitioners and researchers in the field of cloud computing. For practitioners, it is imperative to grasp the nuanced differences and practical applications of various resource management techniques. This understanding can lead to more informed decision-making and the enhancement of cloud infrastructures. For researchers, this study opens new avenues for exploration, particularly in the realm of integrating emerging technologies such as artificial intelligence and machine learning into resource management strategies and future trends. The field of cloud computing is continuously evolving, with technological advancements consistently reshaping the landscape of distributed resource management. This evolution necessitates an ongoing adaptation and innovation in resource management techniques to align with the rapid pace of technological developments.

G. Conclusion

This comprehensive review of distributed resource management in cloud computing has highlighted the complexity and criticality of allocation, scheduling, and provisioning techniques in this domain. The analysis revealed that while there is a plethora of strategies available, each comes with its unique set of advantages and limitations. Allocation techniques are crucial for distributing resources efficiently, scheduling strategies are imperative for maintaining operational efficacy, and provisioning methods play a pivotal role in adapting resource distribution to dynamic demands. The study underscores the importance of a balanced approach in resource management to achieve optimized performance, cost-effectiveness, and scalability in cloud environments.

The implications of this review are significant for both practitioners and researchers. For practitioners, understanding the nuanced differences and applications of various resource management techniques can lead to more informed decisions and improved cloud infrastructures. For researchers, this study opens avenues for further exploration, particularly in the integration of emerging technologies like artificial intelligence and machine learning in resource management strategies. The paper concludes that while significant progress has been made in this field, ongoing advancements in cloud computing technology will continually reshape the landscape of distributed resource management.

H. References

[1] H. Shukur, S. Zeebaree, R. Zebari, O. Ahmed, L. Haji, and D. Abdulqader, "Cache coherence protocols in distributed systems," *Journal of Applied Science and Technology Trends*, vol. 1, no. 3, pp. 92–97, 2020.

[2] S. R. Zeebaree and K. Jacksi, "Effects of processes forcing on CPU and total execution-time using multiprocessor shared memory system," *Int. J. Comput. Eng. Res. Trends*, vol. 2, no. 4, pp. 275–279, 2015.

[3] A. H. Ibrahem and S. R. M. Zeebaree, "Tackling the Challenges of Distributed Data Management in Cloud Computing-A Review of Approaches and Solutions," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 15s, pp. 340–355, 2024.

[4] I. M. I. Zebari, S. R. M. Zeebaree, and H. M. Yasin, "Real time video streaming from multi-source using client-server for video distribution," in *2019 4th Scientific International Conference Najaf (SICN)*, IEEE, 2019, pp. 109–114.

[5] S. R. M. Zeebaree, "DES encryption and decryption algorithm implementation based on FPGA," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 774–781, 2020, doi: 10.11591/ijeecs.v18.i2.pp774-781.

[6] K. H. Sharif, "Client/Servers clustering effects on CPU execution-time, CPU usage and CPU Idle depending on activities of Parallel-Processing-Technique operations".

[7] L. M. Haji, S. R. M. Zeebaree, O. M. Ahmed, M. A. M. Sadeeq, H. M. Shukur, and A. Alkhavvat, "Performance Monitoring for Processes and Threads Execution-Controlling," in *2021 International Conference on Communication & Information Technology (ICICT)*, IEEE, 2021, pp. 161–166.

[8] I. M. Ibrahim, S. R. M. Zeebaree, H. M. Yasin, M. A. M. Sadeeq, H. M. Shukur, and A. Alkhayyat, "Hybrid Client/Server Peer to Peer Multitier Video Streaming," in *2021 International Conference on Advanced Computer Applications (ACA)*, IEEE, 2021, pp. 84–89.

[9] I. S. Abdulkhaleq and S. R. M. Zeebaree, "State of Art for Distributed Databases: Faster Data Access, processing, Growth Facilitation and Improved Communications," *International Journal of Science and Business*, vol. 5, no. 3, pp. 126–136, 2021.

[10] H. M. Zangana and S. R. M. Zeebaree, "Distributed Systems for Artificial Intelligence in Cloud Computing: A Review of AI-Powered Applications and Services," *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, vol. 5, no. 1, pp. 1–20, 2024.

[11] S. R. M. Zeebaree, H. M. Shukur, L. M. Haji, R. R. Zebari, K. Jacksi, and S. M. Abas, "Characteristics and analysis of hadoop distributed systems".

[12] Z. S. Ageed and S. R. M. Zeebaree, "Distributed Systems Meet Cloud Computing: A Review of Convergence and Integration," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 11s, pp. 469–490, 2024.

[13] P. Y. Abdullah, S. R. M. Zeebaree, K. Jacksi, and R. R. Zeabri, "AN HRM SYSTEM FOR SMALL AND MEDIUM ENTERPRISES (SME) S BASED ON CLOUD COMPUTING TECHNOLOGY," 2020.

[14] H. S. Abdullah and S. R. M. Zeebaree, "Distributed Algorithms for Large-Scale Computing in Cloud Environments: A Review of Parallel and Distributed Processing," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 15s, pp. 356–365, 2024.

[15] N. A. Kako, "DDLS: Distributed Deep Learning Systems: A Review," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 10, pp. 7395–7407, 2021.

[16] S. R. M. Zeebaree, R. R. Zebari, K. Jacksi, and D. A. Hasan, "Security Approaches For Integrated Enterprise Systems Performance: A Review".

[17] Y. S. Jghef, S. R. M. Zeebaree, Z. S. Ageed, and H. M. Shukur, "Performance Measurement of Distributed Systems via Single-Host Parallel Requesting using (Single, Multi and Pool) Threads," in *2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA)*, IEEE, 2022, pp. 38–43.

[18] P. Y. Abdullah, H. M. Shukur, and K. Jacksi, "HRM system using cloud computing for Small and Medium Enterprises (SMEs)".

[19] S. R. M. Zeebaree, A. B. Sallow, B. K. Hussan, and S. M. Ali, "Design and Simulation of High-Speed Parallel/Sequential Simplified des Code Breaking Based on FPGA," *2019 International Conference on Advanced Science and Engineering, ICOASE 2019*, pp. 76–81, 2019, doi: 10.1109/ICOASE.2019.8723792.

[20] D. A. Hasan, B. K. Hussan, S. R. M. Zeebaree, D. M. Ahmed, O. S. Kareem, and M. A. M. Sadeeq, "The impact of test case generation methods on the software performance: A review," *International Journal of Science and Business*, vol. 5, no. 6, pp. 33–44, 2021.

[21] R. Jeyaraj, A. Balasubramaniam, A. K. M.A., N. Guizani, and A. Paul, "Resource Management in Cloud and Cloud-Influenced Technologies for Internet of Things Applications," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023, doi: 10.1145/3571729.

[22] H. Malallah *et al.*, "A comprehensive study of kernel (issues and concepts) in different operating systems," *Asian Journal of Research in Computer Science*, vol. 8, no. 3, pp. 16–31, 2021.

[23] M. Saraswat and R. C. Tripathi, "Cloud computing: Analysis of top 5 CSPs in SaaS, PaaS and IaaS platforms," in *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, IEEE, 2020, pp. 300–305.

[24] A. Abid, M. F. Manzoor, M. S. Farooq, U. Farooq, and M. Hussain, "Challenges and Issues of Resource Allocation Techniques in Cloud Computing.," *KSII Transactions on Internet & Information Systems*, vol. 14, no. 7, 2020.

[25] S. H. H. Madni, M. S. A. Latiff, Y. Coulibaly, and S. M. Abdulhamid, "Recent advancements in resource allocation techniques for cloud computing environment: a systematic review," *Cluster Comput*, vol. 20, pp. 2489–2533, 2017.

[26] M. Aldossary, "A Review of Dynamic Resource Management in Cloud Computing Environments.," *Computer Systems Science & Engineering*, vol. 36, no. 3, 2021.

[27] E. Öner and A. H. Özer, "An energy-aware combinatorial auction-based virtual machine scheduling model and heuristics for green cloud computing," *Sustainable Computing: Informatics and Systems*, p. 100889, 2023.

[28] R. Choudhary and S. Perinpanayagam, "Applications of Virtual Machine Using Multi-Objective Optimization Scheduling Algorithm for Improving CPU Utilization and Energy Efficiency in Cloud Computing," *Energies (Basel)*, vol. 15, no. 23, p. 9164, 2022.

[29] U. Jambulingam and K. Balasubadra, "A Unique Multi-Agent-Based Approach for Enhanced QoS Resource Allocation in Multi Cloud Environment while Maintaining Minimized Energy and Maximize Revenue," *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*, vol. 17, no. 2, 2022.

[30] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ML)-centric resource management in cloud computing: A review and future directions," *Journal of Network and Computer Applications*, vol. 204, p. 103405, 2022.

[31] S. Supreeth, K. Patil, S. D. Patil, S. Rohith, Y. Vishwanath, and K. S. Prasad, "An efficient policy-based scheduling and allocation of virtual machines in cloud computing environment," *Journal of Electrical and Computer Engineering*, vol. 2022, 2022.

[32] S. A. A. Matin, S. A. Mansouri, M. Bayat, A. R. Jordehi, and P. Radmehr, "A multi-objective bi-level optimization framework for dynamic maintenance planning of

active distribution networks in the presence of energy storage systems," *J Energy Storage*, vol. 52, p. 104762, 2022.

[33] A. A. Khan, M. Zakarya, and R. Khan, "A Hybrid Heterogeneity Aware Resource Orchestrator for Cloud Platforms," *IEEE Syst J*, vol. 13, no. 4, pp. 3873–3876, 2019.

[34] A. Belgacem, "Dynamic resource allocation in cloud computing: analysis and taxonomies," *Computing*, vol. 104, no. 3, pp. 681–710, 2022.

[35] B. Kruekaew and W. Kimpan, "Multi-objective task scheduling optimization for load balancing in cloud computing environment using hybrid artificial bee colony algorithm with reinforcement learning," *IEEE Access*, vol. 10, pp. 17803–17818, 2022.

[36] S. M. Mohammed, K. Jacksi, and S. Zeebaree, "A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 1, pp. 552–562, 2021.

[37] Y. S. Gan, W. Chen, W.-C. Yau, Z. Zou, S.-T. Liong, and S.-Y. Wang, "3D SOC-Net: Deep 3D reconstruction network based on self-organizing clustering mapping," *Expert Syst Appl*, vol. 213, p. 119209, 2023.

[38] H. Shukur, S. Zeebaree, R. Zebari, D. Zeebaree, O. Ahmed, and A. Salih, "Cloud computing virtualization of resources allocation for distributed systems," *Journal of Applied Science and Technology Trends*, vol. 1, no. 3, pp. 98–105, 2020.

[39] L. M. Haji, S. Zeebaree, O. M. Ahmed, A. B. Sallow, K. Jacksi, and R. R. Zeabri, "Dynamic resource allocation for distributed systems and cloud computing," *TEST Engineering & Management*, vol. 83, no. May/June 2020, pp. 22417–22426, 2020.

[40] F. E. F. Samann, S. R. M. Zeebaree, and S. Askar, "IoT provisioning QoS based on cloud and fog computing," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 29–40, 2021.

[41] M. F. Manzoor, A. Abid, M. S. Farooq, N. A. Nawaz, and U. Farooq, "Resource allocation techniques in cloud computing: A review and future directions," *Elektronika ir Elektrotechnika*, vol. 26, no. 6, pp. 40–51, 2020, doi: 10.5755/j01.eie.26.6.25865.

[42] A. N. Almutlaq and Y. Daadaa, "Auto-scaling approach for cloud based mobile learning applications," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, pp. 472–479, 2019, doi: 10.14569/IJACSA.2019.0100161.

[43] M. A. Jaffar, Q. Ain, and T. S. Choi, "Tumor detection from enhanced magnetic resonance imaging using fuzzy curvelet," *Microsc Res Tech*, vol. 75, no. 4, pp. 499–504, 2012, doi: 10.1002/jemt.21083.

[44] Z. Jian, W. Muqing, and Z. Min, "Joint computation offloading and resource allocation in c-ran with mec based on spectrum efficiency," *IEEE Access*, vol. 7, pp. 79056–79068, 2019, doi: 10.1109/ACCESS.2019.2922702.

[45] S. A. Ali and M. Alam, "Resource-Aware Min-Min (RAMM) algorithm for resource allocation in cloud computing environment," *arXiv preprint arXiv:1803.00045*, 2018.

[46] Z. Li, T. Chu, I. V Kolmanovsky, X. Yin, and X. Yin, "Cloud resource allocation for cloud-based automotive applications," *Mechatronics*, vol. 50, pp. 356–365, 2018.

[47] J. H. Joloudari *et al.*, "Resource allocation optimization using artificial intelligence methods in various computing paradigms: A Review," 2022.

[48] I. M. Ibrahim, "Task scheduling algorithms in cloud computing: A review," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 4, pp. 1041–1053, 2021.

[49] M. Sohani and S. C. Jain, "State-of-the-art survey on cloud computing resource scheduling approaches," in *Ambient Communications and Computer Systems: RACCCS 2017*, Springer, 2018, pp. 629–639.

[50] H. Pallathadka *et al.*, "An investigation of various applications and related challenges in cloud computing," *Mater Today Proc*, vol. 51, pp. 2245–2248, 2022.

[51] M. Ashawa, O. Douglas, J. Osamor, and R. Jackie, "Improving cloud efficiency through optimized resource allocation technique for load balancing using LSTM machine learning algorithm," *Journal of Cloud Computing*, vol. 11, no. 1, p. 87, 2022.

[52] Y.-J. Chiang, Y.-C. Ouyang, A. B. Cremers, and L. Xu, "A load-based scheduling to improve performance in cloud systems," in *2017 First IEEE International Conference on Robotic Computing (IRC)*, IEEE, 2017, pp. 52–59.

[53] R. Kaur, V. Laxmi, and Balkrishan, "Performance evaluation of task scheduling algorithms in virtual cloud environment to minimize makespan," *International Journal of Information Technology*, pp. 1–15, 2022.

[54] J. Logeshwaran, N. Shanmugasundaram, and J. Lloret, "L-RUBI: An efficient load-based resource utilization algorithm for bi-partite scatternet in wireless personal area networks," *International Journal of Communication Systems*, vol. 36, no. 6, p. e5439, 2023.

[55] S. Pal, D. Le, and P. K. Pattnaik, "Virtualization Environment in Cloud Computing," *Cloud Computing Solutions: Architecture, Data Storage, Implementation and Security*, pp. 57–76, 2022.

[56] J. Liu, Y. Zhang, Y. Zhou, D. Zhang, and H. Liu, "Aggressive resource provisioning for ensuring QoS in virtualized environments," *IEEE transactions on cloud computing*, vol. 3, no. 2, pp. 119–131, 2014.

[57] S. Chouliaras and S. Sotiriadis, "An adaptive auto-scaling framework for cloud resource provisioning," *Future Generation Computer Systems*, 2023.

[58] K. Tuli and M. Malhotra, "Optimal Meta-Heuristic Elastic Scheduling (OMES) for VM selection and migration in cloud computing," *Multimed Tools Appl*, pp. 1–27, 2023.

[59] A. Belgacem and K. Beghdad-Bey, "Multi-objective workflow scheduling in cloud computing: trade-off between makespan and cost," *Cluster Comput*, vol. 25, no. 1, pp. 579–595, 2022.

[60] Q. Li, Z. Peng, D. Cui, J. Lin, and J. He, "MHDNNL: A Batch Task Optimization Scheduling Algorithm in Cloud Computing," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 17, no. 1, pp. 1–17, 2022.

[61] E. I. Nehru, J. I. S. Shyni, and R. Balakrishnan, "Auction based dynamic resource allocation in cloud," in *2016 International conference on circuit, power and computing technologies (ICCPCT)*, IEEE, 2016, pp. 1–4.

[62] S. Verma and A. Bala, "Efficient Auto-scaling for Host Load Prediction through VM migration in Cloud," *Concurr Comput*, p. e7925.

[63] B. Sundaravadivazhagan, V. Malathi, and V. Kavitha, "A novel credit grounded job scheduling algorithm for the cloud computing environment," in *2022 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, 2022, pp. 912–919.

[64] M. S. Kumar, A. Choudhary, I. Gupta, and P. K. Jana, "An efficient resource provisioning algorithm for workflow execution in cloud platform," *Cluster Comput*, vol. 25, no. 6, pp. 4233–4255, 2022.
[65] P. Rajasekar and Y. Palanichamy, "A flexible deadline-driven resource provisioning and scheduling algorithm for multiple workflows with VM sharing protocol on WaaS-cloud," *J Supercomput*, pp. 1–31, 2022.
[66] L. Ramachandran, N. C. Narendra, and K. Ponnalagu, "Dynamic provisioning in multi-tenant service clouds," *Service Oriented Computing and Applications*, vol. 6, pp. 283–302, 2012.
[67] K. N. Vhatkar and G. P. Bhole, "Optimal container resource allocation in cloud architecture: A new hybrid model," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 5, pp. 1906–1918, 2022.
[68] R. S. S. Dittakavi, "Evaluating the Efficiency and Limitations of Configuration Strategies in Hybrid Cloud Environments," *International Journal of Intelligent Automation and Computing*, vol. 5, no. 2, pp. 29–45, 2022.
[69] S. Supreeth and K. Patil, "VM Scheduling for Efficient Dynamically Migrated Virtual Machines (VMS-EDMVM) in Cloud Computing Environment.," *KSII Transactions on Internet & Information Systems*, vol. 16, no. 6, 2022.
[70] T. Welsh and E. Benkhelifa, "On resilience in cloud computing: A survey of techniques across the cloud domain," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–36, 2020.
[71] A. Khelifa, T. Hamrouni, R. Mokadem, and F. Ben Charrada, "Combining task scheduling and data replication for SLA compliance and enhancement of provider profit in clouds," *Applied Intelligence*, vol. 51, pp. 7494–7516, 2021.
[72] R. Zolfaghari, A. Sahafi, A. M. Rahmani, and R. Rezaei, "Application of virtual machine consolidation in cloud computing systems," *Sustainable Computing: Informatics and Systems*, vol. 30, p. 100524, 2021.
[73] K. Benhssayen and A. Ettalbi, "Semantic interoperability framework for IAAS resources in multi-cloud environment," *International Journal of Computer Science & Network Security*, vol. 21, no. 2, pp. 1–8, 2021.
[74] S. Soma, "Power Aware Energy Efficient based Virtual Machine Migration Using Enhanced Pelican Remora Optimization in Cloud Center.," *International Journal of Intelligent Engineering & Systems*, vol. 16, no. 6, 2023.
[75] S. R. Swain, A. K. Singh, and C. N. Lee, "Efficient resource management in cloud environment," *arXiv preprint arXiv:2207.12085*, 2022.
[76] S. Chouliaras, "Adaptive resource provisioning in cloud computing environments." Birkbeck, University of London, 2023.
[77] S. Mazumdar and M. Pranzo, "Power efficient server consolidation for cloud data center," *Future Generation Computer Systems*, vol. 70, pp. 4–16, 2017.
[78] G. Han, W. Que, G. Jia, and W. Zhang, "Resource-utilization-aware energy efficient server consolidation algorithm for green computing in IIOT," *Journal of Network and Computer Applications*, vol. 103, pp. 205–214, 2018.
[79] A. Mitra, N. O'Regan, and D. Sarpong, "Cloud resource adaptation: A resource based perspective on value creation for corporate growth," *Technol Forecast Soc Change*, vol. 130, pp. 28–38, 2018.
[80] X. Zhu, J. Wang, H. Guo, D. Zhu, L. T. Yang, and L. Liu, "Fault-tolerant scheduling for real-time scientific workflows with elastic resource provisioning in virtualized

clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 12, pp. 3501–3517, 2016.

[81] G. Jung and K. M. Sim, "Agent-based adaptive resource allocation on the cloud computing environment," in *2011 40th International Conference on Parallel Processing Workshops*, IEEE, 2011, pp. 345–351.

[82] Z.-H. Zhan, X.-F. Liu, Y.-J. Gong, J. Zhang, H. S.-H. Chung, and Y. Li, "Cloud computing resource scheduling and a survey of its evolutionary approaches," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, pp. 1–33, 2015.

[83] A. Barnawi, S. Sakr, W. Xiao, and A. Al-Barakati, "The views, measurements and challenges of elasticity in the cloud: A review," *Comput Commun*, vol. 154, pp. 111–117, 2020.

[84] S. Chowdhury *et al.*, "Unlocking the value of artificial intelligence in human resource management through AI capability framework," *Human Resource Management Review*, vol. 33, no. 1, p. 100899, 2023.

[85] L. He, D. Zou, Z. Zhang, C. Chen, H. Jin, and S. A. Jarvis, "Developing resource consolidation frameworks for moldable virtual machines in clouds," *Future Generation Computer Systems*, vol. 32, pp. 69–81, 2014.

[86] A. R. Hummaida, N. W. Paton, and R. Sakellariou, "Adaptation in cloud resource configuration: a survey," *Journal of Cloud Computing*, vol. 5, pp. 1–16, 2016.

[87] A. Bhattacharya and P. De, "A survey of adaptation techniques in computation offloading," *Journal of Network and Computer Applications*, vol. 78, pp. 97–115, 2017.

[88] S. Di and C.-L. Wang, "Dynamic optimization of multiattribute resource allocation in self-organizing clouds," *IEEE Transactions on parallel and distributed systems*, vol. 24, no. 3, pp. 464–478, 2012.

[89] P. Nehra and N. Kesswani, "Efficient resource allocation and management by using load balanced multi-dimensional bin packing heuristic in cloud data centers," *J Supercomput*, vol. 79, no. 2, pp. 1398–1425, 2023.

[90] W. Shi, L. Zhang, C. Wu, Z. Li, and F. C. M. Lau, "An online auction framework for dynamic resource provisioning in cloud computing," *IEEE/ACM transactions on networking*, vol. 24, no. 4, pp. 2060–2073, 2015.

[91] A. Belgacem, S. Mahmoudi, and M. Kihl, "Intelligent multi-agent reinforcement learning model for resources allocation in cloud computing," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 2391–2404, 2022.

[92] A. M. Alkalbani and F. K. Hussain, "A comparative study and future research directions in cloud service discovery," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, 2016, pp. 1049–1056.

[93] V. P. Nzanzu *et al.*, "Monitoring and resource management taxonomy in interconnected cloud infrastructures: a survey," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 20, no. 2, pp. 279–295, 2022.

[94] U. K. Das, "Resource Scheduling for Infrastructure as a Service (IaaS) in cloud computing." Dublin, National College of Ireland, 2021.

[95] W. Kong, Y. Lei, and J. Ma, "Virtual machine resource scheduling algorithm for cloud computing based on auction mechanism," *Optik (Stuttg)*, vol. 127, no. 12, pp. 5099–5104, 2016.

[96] R. Rashidifar, H. Bouzary, and F. F. Chen, "Resource scheduling in cloud-based manufacturing system: a comprehensive survey," *The International Journal of Advanced Manufacturing Technology*, vol. 122, no. 11–12, pp. 4201–4219, 2022.

[97] F. Ma and Y. Yang, "An energy sentient service brokering strategy in cloud computing," in *2017 29th Chinese Control And Decision Conference (CCDC)*, IEEE, 2017, pp. 4120–4124.

[98] A. Pandey, P. Calyam, Z. Lyu, S. Wang, D. Chemodanov, and T. Joshi, "Knowledge-Engineered Multi-Cloud Resource Brokering for Application Workflow Optimization," *IEEE Transactions on Network and Service Management*, 2022.

[99] S. Potluri and K. S. Rao, "Optimization model for QoS based task scheduling in cloud computing environment," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 1081–1088, 2020.

[100] A. I. Khan, S. A. R. Kazmi, and A. Qasim, "Formal Modeling of Self-Adaptive Resource Scheduling in Cloud.," *Computers, Materials & Continua*, vol. 75, no. 1, 2023.

[101] I. Pietri and R. Sakellariou, "Mapping virtual machines onto physical machines in cloud computing: A survey," *ACM Computing Surveys (CSUR)*, vol. 49, no. 3, pp. 1–30, 2016.

[102] Y. Song, R. Routray, and R. Jain, "Virtual-to-physical mapping inference in virtualized cloud environments," in *2014 IEEE International Conference on Cloud Engineering*, IEEE, 2014, pp. 373–378.

[103] H. Di, V. Anand, H. Yu, L. Li, B. Dong, and Q. Meng, "Reliable virtual infrastructure mapping with efficient resource sharing," in *2013 International Conference on Communications, Circuits and Systems (ICCCAS)*, IEEE, 2013, pp. 198–202.