International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V4I3P112
Eureka Vision Publication | Volume 4, Issue 3, 112-123, 2023

Original Article

Efficient Resource Management and Scheduling in Cloud Computing: A Survey of Methods and Emerging Challenges

Varun Bitkuri[1], Raghuvaran Kendyala[2], Jagan Kurma[3], Jaya Vardhani Mamidala[4], Sunil Jacob Enokkaren[5], Avinash Attipalli[6]
[1]Stratford University ,Software Engineer.
[2]University of Illinois at Springfield, Department of Computer Science.
[3]Christian Brothers University, Computer Information Systems.
[4]University of Central Missouri, Department of Computer Science.
[5]ADP, Solution Architect.
[6]University of Bridgeport, Department of Computer Science.

Abstract - Cloud computing has changed the way modern IT infrastructure works by letting users access computer resources in a way that is scalable, flexible, and on-demand. But making sure that these resources are managed and scheduled well is still important for performance, cost-effectiveness, and energy economy. This essay covers all the different ways to handle resources and make schedules in cloud computing, including old-fashioned methods, smart algorithms, and environmentally friendly ways of doing things. It talks about different ways to schedule and assign resources, such as heuristic, meta-heuristic, machine learning-based, and control-theory-driven models. A lot of attention is paid to managing resources in data centers in a way that saves energy, managing workloads, and optimizing virtual machines in real time. Furthermore, the study identifies key challenges such as workload heterogeneity, elasticity, cost minimization, SLA compliance, and energy-efficient scheduling. The paper also explores architecture models, taxonomy of scheduling layers, and the implications of big data in cloud operations. By consolidating current research and highlighting emerging issues, this paper aims to serve as a valuable reference for researchers and practitioners working toward optimized and sustainable cloud resource management solutions. The insights presented are intended to support the development of scalable frameworks that ensure high availability, reduced latency, and enhanced quality of service in diverse cloud environments.

Keywords - Cloud Computing, Resource Management, scheduling algorithms, Energy Efficiency, Resource Allocation.

1. Introduction

Cloud computing, which offers storage, networking, processing power, and applications through the internet, is an essential component of any contemporary IT architecture [1]. One way these services are distributed is through the SaaS model, another is through the IaaS model, and the third is the PaaS model [2]. IaaS offers consumers computing infrastructure and virtual machines, PaaS provides a platform for developing and deploying applications, and SaaS makes software applications accessible through web interfaces. All these models are typically based on a pay-as-you-use pricing system, providing customers with flexible, location-independent access to computing resources. As depicted in Figure 1, cloud computing evolved from earlier technologies such as grid computing, virtualization, and web-based architectures. These technologies collectively laid the foundation for large-scale distributed infrastructures, where physical machines (PMs) host multiple virtual machines (VMs). Although virtualization improves scalability and resource sharing, many VMs remain powered on even during idle periods, resulting in considerable energy waste [3][4]. Reducing energy usage, boosting system performance, and maintaining service quality have all made efficient management and scheduling of cloud resources critical concerns.

Figure 1. Cloud Computing

Provisioning, allocating, scheduling, and monitoring resources dynamically often in real-time amidst ever-changing workloads is what Cloud Resource Management (CRM) is all about. Effective resource scheduling is important to customer relationship management (CRM) since it determines how tasks are distributed among cloud resources in compliance with SLAs and QoS standards. Since cloud systems are inherently large, unpredictable, and variable, scheduling their resources is an NP-hard problem. Failure to properly allocate resources can result in subpar performance, higher latency, and breaches of service level agreements (SLAs). Due to the complexity of these needs, new and intelligent scheduling methods are urgently needed. The aim of this survey is to thoroughly examine the present status of cloud computing resource scheduling and management. It discusses conventional approaches such as heuristic and meta-heuristic algorithms, as well as advanced methods involving artificial intelligence, machine learning, and energy-aware scheduling. The paper also explores emerging challenges in this domain, including workload prediction accuracy, VM migration overhead, energy efficiency, security, and resource elasticity. A unified understanding of current best practices and identification of knowledge gaps can pave the way for the creation of cloud resource management frameworks that are scalable, energy efficient, and quality of service compliant.

1.1. Structure of the Paper

This paper is organized in the following way: Cloud resource management basics are presented in Section II. Topics covered in Section III include methods and techniques for scheduling. Cloud data centers' energy-efficient resource management is the subject of Section IV. Scheduling and allocating resources are two of the most difficult tasks, as discussed in Section V. The report finishes with a discussion of potential future research topics in Section VII, and Section VI provides a literature assessment of current studies.

2. Fundamentals of Cloud Resource Management

Resource management is essential to every artificial system since it impacts performance, functionality, and cost, the three main metrics used to evaluate systems. Efficient resource management affects performance and cost directly, and indirectly affects the system's functionality because inefficient performance or high costs could cause some functions to be skipped [5]. A cloud is an intricate system that relies on a vast network of shared resources that are vulnerable to requests that are difficult to forecast and to external occurrences over which it has no control. In order to optimize for several objectives at once, cloud resource management necessitates complicated policies. The cloud infrastructure is enormous, and the system's interactions with a large user population are unpredictable, making effective resource management a daunting problem. Because of the sheer number of users and the difficulty in accurately predicting their workload types and intensities, precise global status information is just not achievable. Oversubscribed resources and unruly consumers further complicate resource management [6]. Internal considerations, such as system size, component failure rates, software and hardware heterogeneity, and other factors, influence resource management in addition to external factors.

The rules for managing resources in the cloud can be broadly categorised into five areas: admission control, capacity allocation, load balancing, energy optimisation, and quality of service assurances. The declared goal of an admission control policy is to prevent the system from engaging in actions that conflict with its overall policies. An example of this would be a system that refuses to take on more work if doing so will make it unable to finish contracted or ongoing work [7]. Identity as a service (IaaS), platform as a service (PaaS), software as a service (SaaS), and database as a service (DBaaS) are the four primary cloud delivery types. The assertion of cloud elasticity is challenged in every scenario by the enormous, unpredictable loads that cloud service providers encounter. As an example, web services that experience spikes during certain seasons can have their resources provisioned in advance if the surge can be anticipated. A little more nuanced consideration is required for an unexpected surge.

2.1. Cloud Computing Architecture

The "front end" and the "back end" are the two main components of any cloud architecture. Links to the Internet make the front end, which users may interact with, visible to them. The back end consists of the many cloud service models. See Figure 2 below for a visual representation of the cloud architecture conceptual model as described by NIST [8].

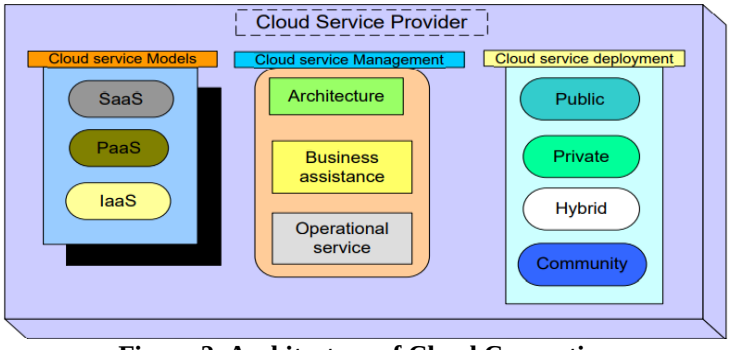

Figure 2. Architecture of Cloud Computing

The cloud computing reference architecture developed by the National Institute of Standards and Technology (NIST) offers a simple yet thorough nomenclature of cloud computing. A brief synopsis of almost every facet of cloud computing is given by the Cloud Computing Architecture. When talking about the architecture, both the front and rear ends are mentioned. Customers make up the front end, whereas service providers make up the rear end. The logistics industry is being approached with a cloud architecture initiative. There is an emphasis on logistics service support through cloud architecture. Offered a multi-tiered approach to logistics cost reduction.

The National Institute of Standards and Technology (NIST) identifies the following characteristics of cloud computing:

- **On-demand self-service:** Interaction with server resources and storage on the network, as well as optimisation when needed, all without the need for a service technician.
- **Broad network access:** Services and functionality are made accessible through Internet connectivity through typical access points, such as cell phones, laptops, and PDAs.
- **Resource pooling:** Offering computer services alongside assistance for clients with several talents is what this provider does best [9]. Resources, both physical and virtual, are allocated and reallocated on the fly to satisfy user demand.
- **Rapid elasticity:** Modifying and implementing features and tools does not take long or necessitate realistic hardware interaction. Additionally, market resources can be scaled up to nearly any extent that is needed.
- **Measured service:** Level of storage, bandwidth, and required user accounts are a few of the variables that influence the configuration and utilisation of data services. There has to be transparency in reporting and verification practices between consumers and providers.

2.2. Types of Cloud Services (IaaS, PaaS, SaaS)

The three primary types of cloud computing service models are: Below, we'll go into more detail about the three service models shown in Figure 3: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS):

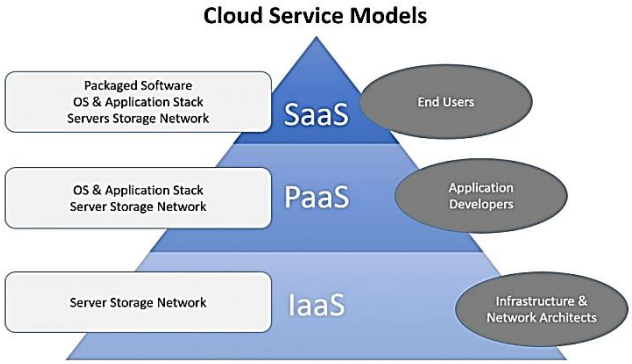

Figure 3. Cloud Service Models (IaaS, PaaS, and SaaS)

2.2.1. Infrastructure as service (IaaS)

IaaS lets users rent computer resources like processing power, memory, and storage space from a service provider and then use those resources to set up and run their own apps. Using virtual computers, users in the IaaS model can access the underlying infrastructure, in contrast to the PaaS model's high level of abstraction. The ability to install any software stack on top of the operating system gives users greater choice with IaaS than with PaaS. Users are liable for operating system updates and patches at the IaaS level, and flexibility is not cheap. S37 and EC2 from Amazon Web Services are two well-known instances of IaaS.

2.2.2. Platform as Service (PaaS)

PAAS allocates and manages the application's storage space, bandwidth, and compute resources. The application's code and implementation are handled by a suite of interconnected tools and services [10]. The responsibility of managing the storage space and data will be handled by the cloud provider. The primary function of PAAS is data security; users can gain access to the software without really buying it.

2.2.3. Software as a service (SaaS)

Cloud service providers run and maintain resources such as application software, operating systems, and more under this model. With the SaaS model, customers get services via the internet and may access them through any web browser. It's like an online application interface. Gmail and Google Docs are just two examples of the hosted applications that are accessible from a wide range of devices. SaaS offers many advantages, one of which is that it removes the customer's burden of buying licenses, installing, upgrading, maintaining, and even using the software on their own computer. Its other advantages include scalability, multitenant efficiency, and configurability. You can see a comparison of cloud services in the table I below, broken down by user count, services offered, and potential applications.

Table 1. Provides The Comparison Of Cloud Services[11]

Service of Cloud	Users	Available Services	Use
IaaS	System Manager	Virtual Machines, OS, Message Queues, Storage, CPU, Memory, Services	Build a system for the handling of service and application development, testing, integration, and deployment.
PaaS	Developers and Deployers	Testing, developing, integrating, and releasing services and applications	Make customer-facing apps and services available to them
SaaS	Business Users	Various online resources such as wikis, blogs, CRM, website testing, and email	Performing desktop-related duties

2.3. Techniques in Cloud Resource Allocation

A systematic strategy, not ad hoc procedures, should underpin cloud resource allocation methodologies. When it comes to managing resources, there are four main ways to put policy into action:

- Control theory: The goal of control theory is to forecast short-term behaviour and ensure system stability through the use of feedback mechanisms. Predicting behaviour on a global scale is not possible with feedback. Oversimplified models have made use of Kalman filters.
- Machine learning. The performance model of the system does not necessary to use machine learning techniques, which is a significant advantage. Multiple autonomic system managers can work together using this method.
- Utility-based. Using a performance model and a method to link user-level performance to cost are prerequisites for utility-based methods.
- Market-oriented mechanisms. Such procedures, like combinatorial auctions for resource bundles, do not necessitate a system model.

3. Scheduling Approaches and Algorithms In Cloud Computing

In a cloud computing setting, resource scheduling can be done at many levels of the service stacks [12]. Based on the architecture that includes IaaS, PaaS, and SaaS stacks, cloud scheduling difficulties can be grouped into three layers: application, virtualisation, and deployment (Figure 4 with explanation following).

- Application layer scheduling is all about efficient and effective resource allocation for software and user applications, jobs, processes, etc., whether those resources are virtual or physical.
- Allocating virtual resources to physical resources in a way that maximises efficiency, minimises waste, and achieves ideal load balancing is the primary goal of virtualisation layer scheduling.
- Worldwide interest is focused on various facets of deployment layer scheduling, such as strategic and optimal infrastructure, outsourcing, service placement, multi-cloud centers, collaborations, data routing, application migration, and more.

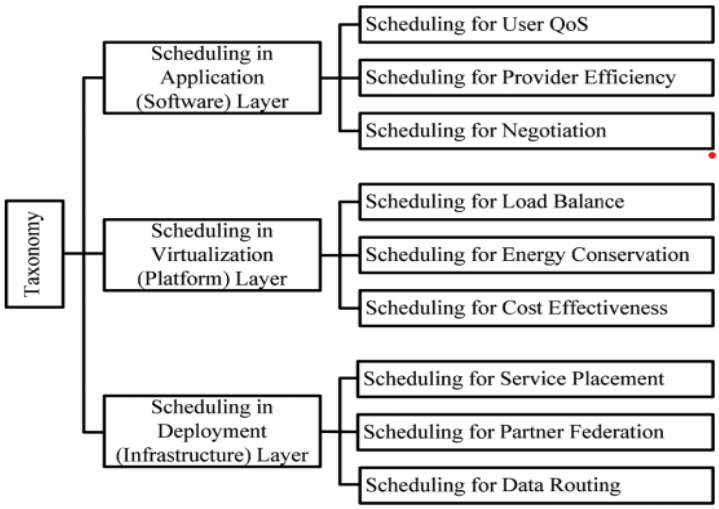

Figure 4. Taxonomy of the Cloud Resource Scheduling

3.1. Scheduling Approaches in Cloud

Different ways of scheduling prepare activities so that they are accessible by resources and aimed at making performance metrics better, like the time needed to complete all the tasks (make span), fair use of resources or balance of work.

3.1.1. Static Scheduling

Static scheduling is based on the idea that efficient hardware utilisation can be achieved when operations begin to share resources without sacrificing performance [13]. When the code contains operations with data and control dependencies that are statically indeterminable or have variable latency, static scheduling plans the beginning times of specific operations to accommodate the worst-case timing scenario. This limits the total throughput and achievable performance.

3.1.2. Dynamic Scheduling

Dynamic scheduling of operations is possible during runtime. By getting beyond the rigidity of conventional static scheduling, it improves throughput in control-dominated and irregular applications. Additional applications that dynamic scheduling can manage include those whose memory accesses are unknown while the application is being built. For instance, in the event that have a statement such as $x[h_1[i]] = g(x[h_2[i]])$, Can begin reading from x after ensuring that there is no read-after-write dependency with any pending store from earlier iterations of the loop. This means that h2 is not the same as any prior or pending store address h1. Once this inequality has been identified, the following operation can start in a dynamically scheduled circuit. If the inequality is not determined, the conflicting memory access will be appropriately stalled.

3.1.3. Workflow-Based Scheduling

The applications in workflow scheduling are depicted as directed acyclic graphs (DAGs), as illustrated in Figure 5. This makes it a generic form of the task scheduling problem. There is a tendency for the terms workflow and DAG to be used interchangeably in papers [14]. Everyone knows that DAG scheduling and all its variations have an NP-complete complexity. And that's in just two basic examples:

- (1) assigning a fixed number of processors jobs with fixed weights.
- (2) NP-completeness was also demonstrated for scheduling jobs with weights of one or two units to two processors.

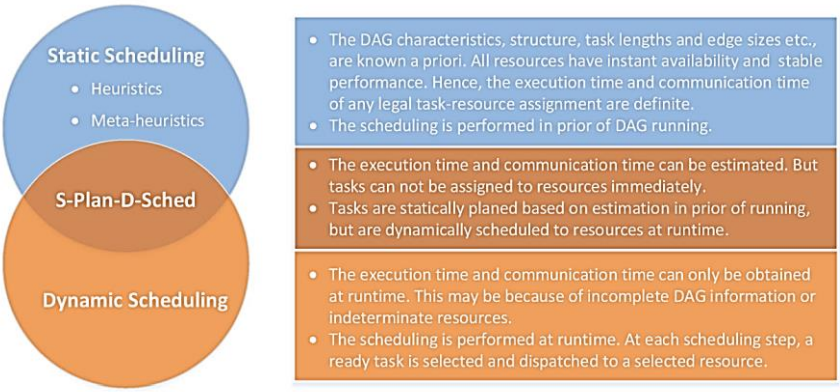

Figure 5. Taxonomy of scheduling workflow

3.2. Scheduling Algorithms in Cloud

Such an algorithm offers guidelines to distribute workloads among available resources (such as virtual machines) for achieving goals like faster completion, balanced use of resources or balanced load.

3.2.1. Minimum Completion Time (MCT)

The MCT scheduling technique takes a set of jobs and assigns them to the VM that can complete them in the least amount of time. This heuristic accomplishes its goal by iteratively scanning each virtual machine (VM) for each task and calculating the time it takes to complete them. The VM-ready time is updated when every job is assigned. With MCT, make span can be reduced, but faster virtual machines will be overloaded while slower ones sit idle. To sum up, MCT reduces makespan but overburdens the faster virtual machines.

3.2.2. Load Balanced Improved Min-Min (LBIMM)

The LBIMM scheduling method is an enhanced variant of the Min-Min method. To guarantee QoS, LIBMM takes into account task priority, which includes execution time and cost. It works wonderfully with smaller datasets but LIBMM doesn't do well with big ones.

3.2.3. Proactive Simulation-Based Scheduling and Load Balancing (PSSLB):

Algorithmically, this method finds out how long it will take for each incoming work to finish on the available virtual machines sorted by decreasing size. A task's row in a completion time matrix represents the time it took for a certain virtual machine to finish using this approach [15]. After selecting the jobs with the shortest completion times, the algorithm deletes this row from the matrix.

4. Energy-Efficient Resource Management in Cloud Data Centers

The cloud is a platform for providing services that consists of distributed software and hardware. It is possible to operate numerous VMs on a single physical machine (PM). Regardless of whether there are no tasks waiting to be processed, virtual machine sets housed on one or more PMs will remain awake [16]. Energy consumption has consequently become an important issue for cloud providers due to the enormous amounts of energy that are wasted.

4.1. Green Computing Based on Artificial Intelligence in Data Centers:

A significant source of unnecessary energy usage in data centers is underutilization. In addition, today's powerful servers can fool the eye into thinking there are numerous smaller VMs, or virtual machines, executing separate applications by utilising virtualisation technology. Data centers may enhance resource utilisation and reduce energy usage with appropriate virtual machine management, which includes allocation, consolidation, and migration [17]. Finding the best possible match between virtual machines (VMs) and servers in terms of resources and enhanced performance metrics is the overarching goal. A pioneering study did something no one had done before: it extended virtualisation systems to include active power management policies that were both rich and effective.

4.2. Effective Resource Management in Cloud Data Center:

Data analytics is a hotspot for research because of the new possibilities and problems it poses for distributed performance management. Researchers are concentrating on the knowledge that can be derived from data collected from system and application performance indicators in order to use these strategies. Various resource management solutions are suggested according to the chosen performance data to be tracked, the method for system abstraction and modelling, and the measures taken to address performance issues [18]. The traditional control loop called the MAPE (Monitor, Analyse, Plan, Execute) loop can be used to explain the major aspects of cloud-based automated resource management. In order to keep tabs on how well the system is running, the constantly gather data on several resource and application attributes [19]. Clean, model, and analyse the data acquired in order to spot any deviations from the system's usual activity. At last, the results of the analysis phase are used to choose the right action, and then the target components are notified to begin executing it.

4.3. Resource Management with Big Data in Cloud Storage:

There are three major levels of resource management in big data cloud environments: setup, operation, and maintenance. Setting up clusters, which can be implemented in on-premise, cloud, or hybrid environments, starts with choosing the right physical or virtual resources (CPU, RAM, storage, network, and increasingly GPUs). After that, a cluster manager is selected to oversee jobs and resources, such as Kubernetes, Docker Swarm, YARN, or Mesos. Depending on whether an application requires batch, stream, or real-time processing, big data frameworks such as Hadoop, Spark, Flink, or Storm are implemented. The operation layer is concerned with meeting service-level agreements (SLAs) by optimizing job scheduling and performance. To forecast resource requirements, performance models are created using either job profiling or historical data. These models can be used to allocate resources statically or dynamically.

The order in which activities are executed is determined by scheduling strategies; FIFO is frequently employed, but it frequently breaks down in complex situations, which has led to the creation of more sophisticated, SLA-aware schedulers. Maintenance Layer: Monitors, detects failures, and scales to guarantee ongoing cluster health. In order to facilitate feedback loops and ML model changes, monitoring tools such as Prometheus and Collectd collect system metrics [20]. Techniques like fault-aware scheduling and work reallocation are used to reduce failures brought on by hardware problems or resource contention. Using cloud elasticity to optimize resource utilization and uphold SLAs, clusters are dynamically scaled up or down to meet workload demands.

5. Challenges in Cloud Scheduling and Resource Allocation

The main difficulties of allocating and scheduling resources in the cloud are summarized in this section. It highlights issues related to workload variability, system scalability, energy efficiency, and the complexity of managing distributed, virtualized environments.

5.1. Predictable and Unpredictable Workloads

Central processing unit, storage, and network resources are mostly virtualised in cloud data centers. When contrasted with traditional data centers, these virtualised resources use far less energy. Users are given a virtualised environment in the form of VMs. Such virtual machines are tasked with handling massive, dynamic workloads.

5.2. Homogenous and Heterogeneous Workloads

One main type of task in the cloud is an application. Those that are very similar to each other in terms of design are in the first group. You can see how much time, memory, storage, and processing power the app needs by looking at that number. The development of cloud systems must take into account the need to allocate both types of workloads.

5.3. Batch Workloads and Transactional Workloads

One major difference between batch and transactional workloads is that batch workloads do not require user input [21]. User input is required for transactional workloads, though. For instance, online transactional systems are a type of transactional workload. Batch workloads do not involve taking proactive measures, in contrast to transactional workloads.

5.4. Elasticity

Cloud elasticity refers to the degree to which it can adapt to changes in resource demand on the fly [22]. The cloud has to be able to automatically recognize when resource demand is about to increase and adjust accordingly. This feature is essential for effective cloud resource management.

5.5. Minimization of Costs and Maximization of Resource Utilization

In cloud resource allocation, minimising total operation cost and optimising resource utilisation are two major criteria that must be satisfied. If a cloud system is really reliable, it will supply its users services with the expectation that they will keep using them [23]. This can only happen if the service provider offers reasonably priced services to its customers.

5.6. VM Migration

One approach to dealing with inadequate cloud resources is virtual machine migration. Host migration allows for the accommodation of virtual machines in terms of resources.

5.7. Handling High Availability for Long-Running Jobs

The execution time of tasks hosted on the cloud is rather long. Consequently, it is critical to have assets ready to go for tasks at all times [24]. The goal is to implement a system that can identify when resources are unavailable or malfunctioning and automatically reallocate them.

5.8. Energy Efficient Allocation

The processing and computing resources housed in cloud data centres are excessively large. Those data centres are going to produce a ton of carbon dioxide.

5.9. Parallel Task Scheduling

The make span of a work could be increased by computing it in parallel. Both independent and dependent tasks exist in the world of work. It is possible to set up many virtual computers to run separate tasks simultaneously [25]. Careful execution is required because of the communication concerns that arise with dependent jobs.

5.10. Networked Cloud

Allocating resources in cloud data centres built within its intradomain has been the subject of numerous techniques and approaches. The algorithms used to allocate resources in more conventional settings would not work in a distributed cloud.

6. Literature Review

This section reviews recent advancements in cloud resource management and scheduling, focusing on emerging techniques like machine learning, blockchain, evolutionary optimization, and advanced algorithms. These methods enhance efficiency, scalability, and cost-effectiveness. A comparative summary is presented in Table II.

Arunarani, Manjula and Sugumaran (2019) provide an in-depth analysis of cloud-friendly work scheduling algorithms together with the relevant metrics. It delves into the numerous challenges and obstacles associated with scheduling approaches. The goal of studying distinct scheduling algorithms is to identify which system attributes should be incorporated and which ones should be ignored. The methodologies, applications, and parameter-based metrics used frameworks provide the framework for the literature survey. They also pinpoint areas that need further investigation into scheduling in the cloud. Scheduling is an essential part of cloud computing since it controls many virtualised resources [26]. Afzal and Kavitha (2019) offer an exhaustive and comprehensive analysis of load balancing methods. To design future effective load balancing algorithms, it is necessary to address critical difficulties while also highlighting the benefits and drawbacks of current systems. New ideas for cloud load balancing are also proposed in the paper. Computing resource performance and efficiency are both negatively impacted by load balancing problems, which are multi-variant and multi-constrained in nature. The two undesirable outcomes of load unbalancing overloading and underloading can be addressed by load balancing solutions [27].

Zhuang and Huang, (2019) the cloud platform resource management in a nutshell. Proper management of the platform's resources is essential for building an efficient cloud computing architecture. they looked into numerous methods for managing and allocating cloud computing resources, such as workload prediction, scheduling, and mapping. I also considered the benefits and drawbacks of each approach. The study lays the groundwork for a new model of managing resources on cloud platforms, with the expectation that it will serve as the basis for the future development of the cloud platform's resource architecture for power trading [28]. Lakkadwala and Kanungo (2018) considered various memory utilisation tactics that aid in assessing and

reducing memory utilisation in a cloud computing setting. On top of that, we're suggesting a brief summary of the notion that they hope will help with scheduling in order to lower memory utilisation in the computational cloud when processing large amounts of data or running processes in the cloud. The promise of efficiently managed resources and effective computing power is central to the idea of cloud computing. Virtualisation makes all these efficiency characteristics possible [29]

Mehmood, Latif and Malik (2018) "Ensemble-based workload prediction mechanism" using stack generalisation as its framework. By comparing it to both the individual and baseline prediction models, the experimental results show how well the suggested model performs. Because the findings of the baseline model were presented as Root Mean Square Error (RMSE), they have utilised this metric for comparison purposes. The suggested method has reduced the root-mean-square error (RMSE) in CPU utilisation prediction by 6% and in memory consumption prediction by 17%. They want to evaluate autonomous learners like K Nearest Neighbour, Decision Tree, Support Vector Machine, Naïve Bayes, and Neural Network in comparison to their proposed ensemble. A 2% improvement in prediction accuracy has been achieved by the suggested ensemble. [30]. Li and Huang (2018) propose a combined approach to resource allocation and job scheduling within the framework of the edge computing paradigm to address the energy efficiency problem of IoT systems.

To be more precise, generalised queueing network models are used to construct the dynamic processes of the IoT services and systems. In order to conduct quantitative studies of performance and energy consumption, these models are used. Markov Decision Process (MDP) is able to find a happy medium between energy costs and quality of service (QoS) needs in the management of resources and the scheduling of tasks. Effective computation, well-managed resources, and their effective utilisation are the goals of cloud computing. Thanks to the idea of virtualisation, all these efficiency features are within reach [31]. Alam, Zulkernine and Haque (2017) In order to provide adequate resources to the consumers, the work's key innovation is to take both reliability and cost into account. Achieving maximum reliability with little expense is our goal with the proposed approach. The purpose of this paper is to offer a heuristic for cloud resource allocation. Several performance studies are presented to support their method, and the results of the simulations demonstrate that their method increases dependability while distributing resources to the consumers [32].

A summary of the literature review is included in Table I, which highlights the main conclusions, difficulties, and suggested future directions of each study.

Table 2. Comparative Summary of Literature Review Based On Resource Management and Scheduling In Cloud Computing

Reference	Study On	Approach	Key Findings	Challenges	Future Direction
Arunarani, et.al. (2019)	An exhaustive analysis of cloud computing work scheduling algorithms and metrics	Classification based on: Scheduling methods (heuristic, metaheuristic, hybrid) Applications (QoS-based, energy-aware, cost-effective) Parameter-based scheduling (deadline, cost, makespan, etc.)	Identification of effective scheduling strategies for different cloud scenarios Analysis of performance-affecting parameters Emphasis on role of virtualization and dynamic resource allocation	Complexity in selecting optimal scheduling technique Trade-offs between QoS parameters Scalability and heterogeneity in resource availability	Development of adaptive and intelligent scheduling algorithms Integration with emerging paradigms like fog and edge computing • Real-time scheduling using AI/ML approaches
Afzal et.al. (2019)	Approaches to cloud-based load balancing	Categorization and evaluation of: Static and dynamic techniques Centralized and distributed mechanisms Heuristic-based and hybrid models	Load imbalance leads to resource under-utilization or over-utilization Effective load balancing improves performance and reduces SLA violations, and requires intelligent algorithms to manage cloud dynamics	Multi-objective nature of the problem •Handling resource heterogeneity and task diversity Latency in load redistribution	Use of machine learning and soft computing for predictive load balancing Real-time, decentralized load balancing, Energy-aware and context-driven balancing algorithms
Zhuang & Huang (2019)	Resource allocation and management in cloud	Comparative study of workload prediction, scheduling, mapping	Comprehensive analysis of methods with pros and cons	Lack of unified framework, varied effectiveness per scenario	Propose new architecture for cloud resource management in

	platforms				power trading
Lakkadwala & Kanungo (2018)	Memory utilization in cloud computing	Survey of memory usage reduction strategies	Memory-aware scheduling reduces impact of data-intensive processes	Memory bottlenecks during high transfer and execution	Propose optimized scheduling based on memory usage trends
Mehmood, Latif & Malik (2018)	Workload prediction for resource provisioning	Ensemble learning via stacked generalization	6–17% RMSE reduction, ~2% accuracy gain over individual models	Handling variability in resource demands	Enhance prediction mechanisms and integrate into auto-scaling
Li & Huang (2018)	Energy-efficient IoT systems under edge computing	MDP-based resource allocation and task scheduling	Balanced energy cost and QoS, performance-energy tradeoff	Modeling dynamic IoT processes with queue networks	Refine MDP modeling and extend to diverse IoT applications
Alam, Zulkernine & Haque (2017)	Cost-reliable cloud resource allocation	Heuristic maximizing reliability and minimizing cost	Improved reliability with cost-efficiency in simulations	Tradeoff between cost and performance reliability	Improve heuristic for real-world deployment and larger-scale validation

7. Conclusion and Future Work

Cloud computing performance, scalability, energy efficiency, and cost-effectiveness are dependent on efficient scheduling and management of resources. Several strategies have been investigated in this survey for optimising scheduling and allocation of cloud resources, including control theory, machine learning, utility-based approaches, and market-oriented processes. It went on to discuss new approaches to dealing with the difficulties of heterogeneous and ever-changing cloud environments, as well as energy-efficient techniques and workload-aware strategies. The paper further highlighted the complexities involved in workload distribution, VM migration, and SLA compliance. By identifying the limitations of current approaches and presenting a comparative overview, this study provides a foundational understanding for researchers and practitioners.

Future research should focus on developing adaptive, AI-driven resource management frameworks capable of handling real-time decision-making under highly dynamic cloud workloads. Integration of multi-cloud and edge computing paradigms with existing cloud infrastructures presents promising opportunities to enhance resource elasticity and responsiveness. Additionally, there is a need to investigate more robust and lightweight security-aware scheduling algorithms, energy-aware service placement, and fault-tolerant resource provisioning to ensure reliability and sustainability. Exploring federated learning and decentralized architectures can further enable more resilient and privacy-preserving cloud systems

References

[1] M. Kaur and H. Singh, “A review of cloud computing security issues,” *Int. J. Grid Distrib. Comput.*, vol. 8, no. 5, pp. 215–222, 2015, doi: 10.14257/ijgdc.2015.8.5.21.

[2] A. Agarwal and S. Jain, “Efficient Optimal Algorithm of Task Scheduling in Cloud Computing Environment,” *Int. J. Comput. Trends Technol.*, vol. 9, no. 7, 2014, doi: 10.14445/22312803/ijctt-v9p163.

[3] B. Jennings and R. Stadler, “Resource Management in Clouds: Survey and Research Challenges,” *J. Netw. Syst. Manag.*, vol. 23, no. 3, pp. 567–619, Jul. 2015, doi: 10.1007/s10922-014-9307-7.

[4] S. Garg, “Predictive Analytics and Auto Remediation using Artificial Intelligence and Machine learning in Cloud Computing Operations,” *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 7, no. 2, 2019.

[5] J. K. Konjaang, J. Y. Maipan-uku, and K. Kennedy, “An Efficient Max-Min Resource Allocator and Task Scheduling Algorithm in Cloud Computing Environment,” *Int. J. Comput. Appl.*, vol. 142, no. 8, pp. 25–30, May 2016, doi: 10.5120/ijca2016909884.

[6] A. Kushwaha, P. Pathak, and S. Gupta, “Review of optimize load balancing algorithms in cloud,” *Int. J. Distrib. Cloud Comput.*, vol. 4, no. 2, pp. 1–9, 2016.

[7] S. Mustafa, B. Nazir, A. Hayat, A. U. R. Khan, and S. A. Madani, “Resource management in cloud computing: Taxonomy, prospects, and challenges,” *Comput. Electr. Eng.*, vol. 47, pp. 186–203, 2015, doi: 10.1016/j.compeleceng.2015.07.021.

[8] N. U. Saqib, M. Arora, and S. Chopra, “Cloud Computing Architecture Issues and Future Research Directions,” vol. 5, no. 11, pp. 532–537, 2018.

[9] I. Odun-Ayo, M. Ananya, F. Agono, and R. Goddy-Worlu, “Cloud Computing Architecture: A Critical Analysis,” in *Proceedings of the 2018 18th International Conference on Computational Science and Its Applications, ICCSA 2018*, 2018. doi: 10.1109/ICCSA.2018.8439638.

[10] S. V. Mohan and S. S. Sathyanathan, “Research in Cloud Computing-An Overview,” *Int. J. Distrib. Cloud Comput.*, 2015,

doi: 10.21863/ijdcc/2015.3.1.002.
[11] H. Mehta, V. K. Prasad, and M. Bhavsar, "Efficient Resource Scheduling in Cloud Computing," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 3, pp. 809–815, 2017, doi: 10.26483/ijarcs.v8i3.3104.
[12] 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 Comput. Surv.*, vol. 47, no. 4, Jul. 2015, doi: 10.1145/2788397.
[13] P. Pathak, A. Shrivastava, and S. Gupta, "A Survey on Various Security Issues in Delay Tolerant Networks," *J. Adv. Shell Program.*, vol. 2, no. 2, pp. 12–18, 2015.
[14] F. Wu, Q. Wu, and Y. Tan, "Workflow scheduling in cloud: a survey," *J. Supercomput.*, vol. 71, no. 9, pp. 3373–3418, 2015, doi: 10.1007/s11227-015-1438-4.
[15] S. S. S. Neeli, "Serverless Databases: A Cost-Effective and Scalable Solution," *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 7, no. 6, p. 7, 2019.
[16] W. Zhao, X. Wang, S. Jin, W. Yue, and Y. Takahashi, "An energy efficient task scheduling strategy in a cloud computing system and its performance evaluation using a two-dimensional continuous time markov chain model," *Electron.*, 2019, doi: 10.3390/electronics8070775.
[17] X. Jin, F. Zhang, A. V. Vasilakos, and Z. Liu, "Green Data Centers: A Survey, Perspectives, and Future Directions," 2016.
[18] S. K. Moghaddam, R. Buyya, and K. Ramamohanarao, "Performance-aware management of cloud resources: A taxonomy and future directions," *ACM Comput. Surv.*, 2019, doi: 10.1145/3337956.
[19] A. J. Younge, G. von Laszewski, L. Wang, S. Lopez-Alarcon, and W. Carithers, "Efficient resource management for Cloud computing environments," in *International Conference on Green Computing*, IEEE, Aug. 2010, pp. 357–364. doi: 10.1109/GREENCOMP.2010.5598294.
[20] M. T. Islam and R. Buyya, "Resource Management and Scheduling for Big Data Applications in Cloud Computing Environments," 2019. doi: 10.4018/978-1-5225-8407-0.ch001.
[21] A. Khajeh-Hosseini, I. Sommerville, and I. Sriram, "Research Challenges for Enterprise Cloud Computing," 2010.
[22] S. Di and C.-L. Wang, "Dynamic Optimization of Multiattribute Resource Allocation in Self-Organizing Clouds," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 3, pp. 464–478, Mar. 2013, doi: 10.1109/TPDS.2012.144.
[23] A. M. H. Kuo, "Opportunities and challenges of cloud computing to improve health care services," 2011. doi: 10.2196/jmir.1867.
[24] A. E. Evwiekpaefe and F. Ajakaiye, "The Trend and Challenges of Cloud Computing: A Literature Review," *Acad. J. Interdiscip. Stud.*, 2013, doi: 10.5901/ajis.2013.v2n10p9.
[25] C. Papagianni, A. Leivadeas, S. Papavassiliou, V. Maglaris, C. Cervello-Pastor, and A. Monje, "On the optimal allocation of virtual resources in cloud computing networks," *IEEE Trans. Comput.*, vol. 62, no. 6, pp. 1060–1071, Jun. 2013, doi: 10.1109/TC.2013.31.
[26] A. R. Arunarani, D. Manjula, and V. Sugumaran, "Task scheduling techniques in cloud computing: A literature survey," *Futur. Gener. Comput. Syst.*, 2019, doi: 10.1016/j.future.2018.09.014.
[27] S. Afzal and G. Kavitha, "Load balancing in cloud computing – A hierarchical taxonomical classification," *J. Cloud Comput.*, vol. 8, no. 1, p. 22, Dec. 2019, doi: 10.1186/s13677-019-0146-7.
[28] W. Zhuang and L. Huang, "Overview of cloud computing resource allocation and management technology," in *2019 6th International Conference on Systems and Informatics, ICSAI 2019*, 2019. doi: 10.1109/ICSAI48974.2019.9010101.
[29] P. Lakkadwala and P. Kanungo, "Memory utilization techniques for cloud resource management in cloud computing environment: A survey," in *2018 4th International Conference on Computing Communication and Automation, ICCCA 2018*, 2018. doi: 10.1109/CCAA.2018.8777457.
[30] T. Mehmood, S. Latif, and S. Malik, "Prediction of Cloud Computing Resource Utilization," in *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*, IEEE, Oct. 2018, pp. 38–42. doi: 10.1109/HONET.2018.8551339.
[31] S. Li and J. Huang, "Energy efficient resource management and task scheduling for IoT services in edge computing paradigm," in *Proceedings - 15th IEEE International Symposium on Parallel and Distributed Processing with Applications and 16th IEEE International Conference on Ubiquitous Computing and Communications, ISPA/IUCC 2017*, 2018. doi: 10.1109/ISPA/IUCC.2017.00129.
[32] A. B. M. B. Alam, M. Zulkernine, and A. Haque, "A Reliability-Based Resource Allocation Approach for Cloud Computing," in *2017 IEEE 7th International Symposium on Cloud and Service Computing (SC2)*, 2017, pp. 249–252. doi: 10.1109/SC2.2017.46.
[33] Rajiv, C., Mukund Sai, V. T., Venkataswamy Naidu, G., Sriram, P., & Mitra, P. (2022). Leveraging Big Datasets for Machine Learning-Based Anomaly Detection in Cybersecurity Network Traffic. *J Contemp Edu Theo Artific Intel: JCETAI/102*.
[34] [34]Sandeep Kumar, C., Srikanth Reddy, V., Ram Mohan, P., Bhavana, K., & Ajay Babu, K. (2022). Efficient Machine Learning Approaches for Intrusion Identification of DDoS Attacks in Cloud Networks. *J Contemp Edu Theo Artific Intel: JCETAI/101*.

[35] Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2020). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

[36] Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

[37] Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2022). Designing an Intelligent Cybersecurity Intrusion Identify Framework Using Advanced Machine Learning Models in Cloud Computing. *Universal Library of Engineering Technology*, (Issue).

[38] Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2022). Leveraging Artificial Intelligence Algorithms for Risk Prediction in Life Insurance Service Industry. *Available at SSRN 5459694*.

[39] Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(3), 70-80.

[40] Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. S. V. Efficient Framework for Forecasting Auto Insurance Claims Utilizing Machine Learning Based Data-Driven Methodologies. *International Research Journal of Economics and Management Studies IRJEMS*, *1*(2).

[41] Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. (2022). Blockchain Technology in Supply Chain and Logistics: A Comprehensive Review of Applications, Challenges, and Innovations. *International Journal of Emerging Research in Engineering and Technology*, *3*(3), 99-107.

[42] Narra, B., Vattikonda, N., Gupta, A. K., Buddula, D. V. K. R., Patchipulusu, H. H. S., & Polu, A. R. (2022). Revolutionizing Marketing Analytics: A Data-Driven Machine Learning Framework for Churn Prediction. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(2), 112-121.

[43] Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., & Gupta, A. K. BLOCKCHAIN TECHNOLOGY AS A TOOL FOR CYBERSECURITY: STRENGTHS, WEAKNESSES, AND POTENTIAL APPLICATIONS.

[44] Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2022). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

[45] Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

[46] HK, K. (2020). Design of Efficient FSM Based 3D Network on Chip Architecture. *INTERNATIONAL JOURNAL OF ENGINEERING*, *68*(10), 67-73.

[47] Krutthika, H. K. (2019, October). Modeling of Data Delivery Modes of Next Generation SOC-NOC Router. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-6). IEEE.

[48] Ajay, S., Satya Sai Krishna Mohan G, Rao, S. S., Shaunak, S. B., Krutthika, H. K., Ananda, Y. R., & Jose, J. (2018). Source Hotspot Management in a Mesh Network on Chip. In *VDAT* (pp. 619-630).

[49] Nair, T. R., & Krutthika, H. K. (2010). An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System. *arXiv preprint arXiv:1001.3781*.

[50] Gopalakrishnan Nair, T. R., & Krutthika, H. K. (2010). An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System. *arXiv e-prints*, arXiv-1001.

[51] Krutthika H. K. & A.R. Aswatha. (2021). Implementation and analysis of congestion prevention and fault tolerance in network on chip. *Journal of Tianjin University Science and Technology*, *54*(11), 213–231. https://doi.org/10.5281/zenodo.5746712

[52] Krutthika H. K. & A.R. Aswatha. (2020). FPGA-based design and architecture of network-on-chip router for efficient data propagation. *IIOAB Journal*, *11*(S2), 7–25.

[53] Krutthika H. K. & A.R. Aswatha (2020). Design of efficient FSM-based 3D network-on-chip architecture. *International Journal of Engineering Trends and Technology*, *68*(10), 67–73. https://doi.org/10.14445/22315381/IJETT-V68I10P212

[54] Krutthika H. K. & Rajashekhara R. (2019). Network-on-chip: A survey on router design and algorithms. *International Journal of Recent Technology and Engineering*, *7*(6), 1687–1691. https://doi.org/10.35940/ijrte.F2131.037619

[55] Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models. *International Journal of AI, BigData, Computational and Management Studies*, *2*(2), 55-65.

[56] Gangineni, V. N., Tyagadurgam, M. S. V., Chalasani, R., Bhumireddy, J. R., & Penmetsa, M. (2021). Strengthening Cybersecurity Governance: The Impact of Firewalls on Risk Management. *International Journal of AI, BigData, Computational and Management Studies*, *2*, 10-63282.

[57] Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Gangineni, V. N. (2021). An Advanced Machine Learning Models Design for Fraud Identification in Healthcare Insurance. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 26-34.

[58] Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., & Polam, R. M. (2021). Advanced Machine Learning Models for Detecting and Classifying Financial Fraud in Big Data-Driven. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(3), 39-46.

[59] Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2021). Enhancing IoT (Internet of Things) Security through Intelligent Intrusion Detection Using ML Models. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 27-36.

[60] Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2021). Smart Healthcare: Machine Learning-Based Classification of Epileptic Seizure Disease Using EEG Signal Analysis. *International Journal of Emerging Research in Engineering and Technology*, *2*(3), 61-70.

[61] Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2021). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(2), 26-34.

[62] Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., & Pabbineedi, S. (2021). Next-Generation Cybersecurity: The Role of AI and Quantum Computing in Threat Detection. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(4), 54-61.

[63] Polu, A. R., Vattikonda, N., Gupta, A., Patchipulusu, H., Buddula, D. V. K. R., & Narra, B. (2021). Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques. *Available at SSRN 5297803*.

[64] Kalla, D. (2022). AI-Powered Driver Behavior Analysis and Accident Prevention Systems for Advanced Driver Assistance. *International Journal of Scientific Research and Modern Technology (IJSRMT) Volume*, *1*.

[65] Dinesh, K. (2022). Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. *International Advanced Research Journal in Science, Engineering and Technology*.

[66] Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. *International Journal of Computing and Artificial Intelligence*, *2*(2), 55-62.

[67] Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. *ESP Journal of Engineering & Technology Advancements (ESP-JETA)*, *1*(1), 150-157.

[68] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2021). Facial Emotion and Sentiment Detection Using Convolutional Neural Network. *Indian Journal of Artificial Intelligence Research (INDJAIR)*, *1*(1), 1-13.

[69] Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. *Available at SSRN 5266517*.