Intelligent VM Scheduling in Cloud Computing: Survey of Existing Methods and Proposal of a New Hybrid Heuristic-ML Framework

Chaimae BAHIJ [1], Mohamed EL GHMARY[2], Hassan ECHOUKAIRI[1]

[1] Computer Sciences Department, Faculty of Sciences, Mohammed V University in Rabat, Morocco
[2] Computer Sciences Department, FSDM, Sidi Mohamed Ben Abdellah University, Morocco
[1] Computer Sciences Department, Faculty of Sciences, Mohammed V University in Rabat, Morocco

Abstract. This paper explores sophisticated Virtual Machine (VM) scheduling approaches in cloud computing and their significance to enhance resource distribution, improve system efficiency, and cost reduction. It provides a recent overviews of key scheduling algorithms, including heuristic, metaheuristic, advanced machine learning-based and hybrid approaches, while assessing their respective strengths, weaknesses and practical applications. The discussion encompasses their applications in managing workloads, optimizing costs, enhancing energy efficiency, improving Quality of Service (QoS) and with a particular focus on scalability and real-time scheduling in cloud settings. Furthermore, the paper analyzes scheduling strategies adopted by major cloud providers through real-world case studies. Ultimately, our analysis identifies critical VM scheduling trade-offs, provides optimization guidelines, and validates the efficacy of hybrid adaptive methods via a new proposed heuristic–machine learning model for dynamic cloud environments.

Keywords: Virtual Machine (VM) Scheduling, Optimization, Resource Allocation, Cloud Computing.

1 INTRODUCTION

Virtual Machine (VM) scheduling is a critical operation in cloud computing (CC) for optimal allocation and use of resources in a cloud environment. It involves virtual machine allocation to physical servers with such performance optimization, execution time reduction, and system workloads balancing as feasible. As the usage of cloud environments increases for bulk computing, VM scheduling is critical to ensuring efficiency as well as reducing operational expenses. Cloud computing relies on virtualization, which supports various users sharing underlying physical resources securely and efficiently. Inefficient scheduling of VMs can lead to resource underutilization, system overload, and excess energy consumption. Performance is influenced by scheduling by controlling the resource allocation, load balancing, and fault tolerance. Inefficient scheduling can lead to delayed execution, wasteful energy consumption, and excess cost owing to poor utilization of computer power.VM scheduling may be classified as static and dynamic methods. Static scheduling statically allocates resources based on set parameters and is suited for uniform workloads, whereas dynamic scheduling continuously re-allocates resources based on real demand. Also, heuristic, real-time, and workflow-based scheduling methods offer optimization mechanisms tailored to specific computing needs.
Effective VM scheduling enhances cloud efficiency by maximizing resource utilization, service availability, and infrastructure performance. Scheduling methods are used by cloud providers to manage workload distribution, reduce costs, and maintain quality of service (QoS). The scheduling algorithms must also address traffic load balancing, energy utilization, and system failures. Fault-tolerant techniques such as dynamic resource scheduling avoid task failure and performance fluctuations [1]. In the following sections, we present a detailed review of VM scheduling algorithms categorized by heuristics, metaheuristics, machine learning, rule-based, and hybrid approaches. We then formalize the VM scheduling process and propose a mathematical model, followed by a new hybrid intelligent framework that combines heuristics, machine learning, and automation. Real-world case studies from major cloud providers are analyzed, and the main performance metrics are evaluated. Finally, we discuss the main challenges in VM scheduling, explore recent trends and future directions, and conclude with the significance of adaptive and automated scheduling solutions in modern cloud environments.

2 BACKGROUND AND FUNDAMENTAL CONCEPTS

2.1 Cloud Computing Basics

Cloud computing is an innovative paradigm that provides on-demand network access to computing resources such as servers, storage, networking, and software over the internet, allowing organizations to dynamically scale infrastructure and reduce costs. It is generally provided in three models of service: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [2]. Scalability, flexibility, cost-effectiveness, high availability, and disaster recovery are the major characteristics . Cloud computing has its foundation in virtualization, enabling virtual replicas of physical resources so that several systems and applications share the same hardware. This enables resource pooling, load balancing, and dynamic scaling while improving security and performance . By leveraging virtualization, cloud computing is more flexible and efficient, complemented by modern technologies such as containerization (e.g., Docker, Kubernetes) and serverless computing .

2.2 Scheduling Basics

Cloud scheduling is the deliberate allocation of workloads to available resources with the aim of driving performance, utilization, and cost-effectiveness [3]. Scheduling in the traditional sense relies on static resources, whereas cloud scheduling must manage dynamic environments, ensure Quality of Service (QoS) guarantees, and address constraints such as power consumption and resource availability in distributed data centers [4]. Effective VM scheduling leads to balanced loads, prevents system overloading, enables fault tolerance, and addresses dynamic demands using hybrid methods like Constraint Programming (CP) and Operations Research (OR). Scheduling algorithms are compared based on their efficiency, which is computed using performance metrics like makespan, resource utilization, load balancing, energy efficiency, execution cost, QoS, fault tolerance, and scalability [1]. These metrics guide the development of advanced scheduling capabilities that ensure healthy, affordable, and high-performing cloud data centers.

3 VM SCHEDULING IN CLOUD COMPUTING: ALGORITHMIC APPROACHES, PROCESSES, AND FORMAL MODELING

VM scheduling plays a crucial role in cloud computing by ensuring the efficient allocation of resources to incoming tasks, directly impacting performance and cost. This section presents a classification of common scheduling approaches, ranging from heuristics to hybrid and ML-based methods, followed by a description of the task-to-VM and VM-to-host mapping process. It concludes with a formal mathematical model framing VM scheduling as a multi-objective optimization problem, which forms the basis for intelligent scheduling strategies.

3.1 Classification of Algorithms

Virtual machine scheduling algorithms can be categorized into several classes. Heuristic algorithms such as Genetic Algorithms (GA) and Ant Colony Optimization (ACO) provide good-enough solutions using evolutionary and bio-inspired mechanisms, suitable for exploring large solution spaces efficiently [5]. Metaheuristics like Particle Swarm Optimization (PSO) and Simulated Annealing (SA) generalize these ideas to search globally and avoid local optima through collaborative or probabilistic strategies[6]. Machine learning-based methods including reinforcement learning and deep learning—learn adaptive scheduling policies from data, enabling better decisions in dynamic cloud environments [7]. Rule-based and priority-based techniques allocate resources based on fixed rules or urgency criteria, offering simplicity at the cost of flexibility [8]. Lastly**,** hybrid approaches combine these techniques (e.g., heuristics with ML) to leverage their respective strengths in handling complex and dynamic scheduling contexts [9].The conceptual comparison of scheduling algorithms described above is graphically presented in Figure 1:

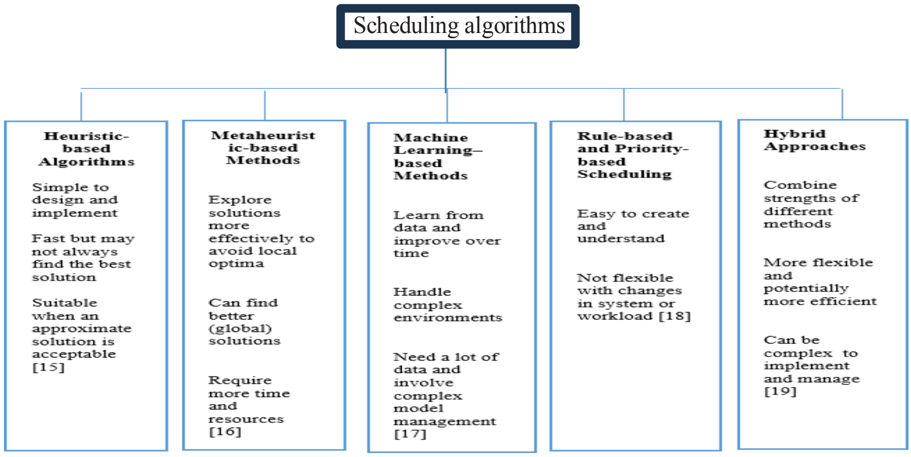

Fig. 1. Scheduling approaches

In addition to the descriptions above, Table 1 provides a summarized comparative view of the main categories of VM scheduling strategies in terms of their type, advantages, and limitations.

Table 1. Summary comparison of VM scheduling strategies.

Approach	Type	Strengths	Limitations
Heuristic-based	Heuristic	Simple, fast	Not always optimal
Metaheuristic-based	Metaheuristic	Avoids local optima, better solutions	Requires more computation
ML-based	Intelligent	Adaptive, learns from data	large datasets, complex to tune
Rule/Priority-based	Deterministic	Easy to implement, transparent	Inflexible in dynamic environments
Hybrid	Combined	Combines strengths of multiple methods	More complex to design and validate

3.2 VM scheduling process

The VM scheduling process in cloud computing typically consists of two essential mapping stages: first, allocating tasks or applications to appropriate virtual machines (task-to-VM mapping); second, assigning these VMs to suitable physical hosts (VM-to-host mapping). This process aims to optimize several objectives such as execution time, energy consumption, and cost, while satisfying QoS constraints and resource availability [10] .

During the task-to-VM phase, each job is evaluated based on resource requirements (CPU, memory, I/O) and is mapped to a VM that meets these needs. Then, in the VM-to-host mapping phase, the cloud infrastructure selects the most appropriate host based on current load, capacity, and power efficiency. These mappings are guided by scheduling algorithms and refined using adaptive methods such as machine learning for dynamic optimization. This process is depicted in figure 2 below:

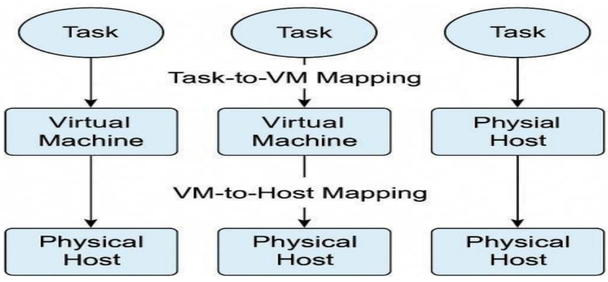

Fig. 2. Scheduling process

3.3 Mathematical Formulation of the Scheduling Problem

The mathematical model is inspired by multi-objective scheduling formulations as presented by Xu and Li [10]. VM scheduling in cloud computing can be formally defined as a multi-objective optimization problem. The goal is to assign a set of tasks to virtual machines and subsequently map those VMs to physical hosts in a way that optimizes certain performance metrics, such as execution cost, energy consumption, and resource utilization, while satisfying system constraints.

Let T={ $t_1, t_2 \ldots, t_n$ } be the set of tasks to be scheduled, V={ $v_1, v_2 \ldots, v_m$ } the set of available virtual machines, and H={ $h_1, h_2 \ldots, h_k$ } the set of physical hosts. Let c_{ij} denote the cost of assigning task t_i to VM v_j, E_{jk} the energy consumed by placing VM v_j on host h_k, and U_{jk} the resource utilization of VM v_j on host h_k. Let binary variables $x_{ij} \in \{0,1\}$ indicate whether task t_i is assigned to VM v_j, and $y_{jk} \in \{0,1\}$ indicate whether VM v_j is hosted on physical machine h_k. The multi objective formulation consists of the following goals:

Minimize total execution cost: $\min \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}.C_{ij}$ (1)

Minimize total energy consumption: $\min \sum_{j=1}^{m}\sum_{k=1}^{k} y_{jk}.E_{jk}$ (2)

Maximize resource utilization: $\max \sum_{j=1}^{m}\sum_{k=1}^{k} y_{jk}.U_{jk}$ (3)

But the scheduling process must satisfy the following system constraints:

Each task is assigned to exactly one VM: $\sum_{j=1}^{m} x_{ij} = 1, \forall i \in \{1,...,n\}$ (4)

VM resource limits are not exceeded: $v_j h_k$ (5)

Host resource capacity must not be exceeded:

$$\sum_{j=1}^{m} y_{jk}.R_j^{vm} \le R_k^{Host}, \forall k \in \{1,...,k\} \tag{6}$$

Where R_i is the resource demand of task t_i, R_j^{VM} is the resource capacity of VM v_j, and R_k^{Host} is the capacity of physical host h_k.

This formulation provides a solid foundation for implementing scheduling strategies using heuristic, metaheuristic, or machine learning-based optimization techniques. It also allows for evaluating the trade-offs between execution cost, energy efficiency, and resource utilization under real-world cloud infrastructure constraints. The flowchart in Figure 3 illustrates the interaction between tasks, virtual machines (VMs), and hosts.

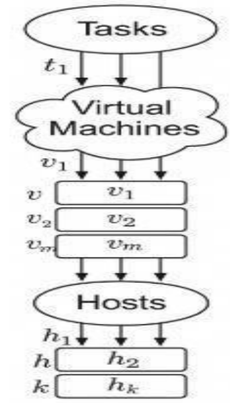

Fig. 3. Mathematical formulation of the scheduling problem [10].

3.4 Evaluation Metrics

To quantify the performance of VM scheduling algorithms, several important performance parameters are employed: Resource utilization computes how efficiently the computing resources (CPU, memory, bandwidth) are utilized. High resource utilization indicates efficient utilization of resources with little idle time . Makespan is the time taken to complete a set of tasks. Makespan minimization is critical for system throughput and response enhancement . Cost compares the cost of operation incurred due to usage of resources. Scheduling is supposed to be cost-effective in nature but deliver as per the needs . Energy Efficiency studies energy usage on utilization of the tasks. Energy-efficient scheduling saves in operation cost and even saves the environment . By the understanding and application of these classifications, comparisons, and evaluation criteria, system administrators and cloud providers are able to select and tailor VM scheduling algorithms best suited to their own specific operational goals and limitations.

4 APPLICATIONS OF ADVANCED VM SCHEDULING

VM scheduling supports several key objectives in cloud computing. It helps manage both static and dynamic workloads, ensuring efficient resource allocation even under fluctuating demand (e.g., Google's dynamic scheduler) [10]. It plays a central role in cost optimization through automated VM management and dynamic scaling to prevent overprovisioning . From an environmental perspective, energy-efficient scheduling reduces power consumption via workload consolidation and the use of low-energy servers [11]. Additionally, Quality of Service (QoS) is enhanced by improving reliability, scalability, and response time guarantees for latency-sensitive tasks [12]. As illustrated in Figure 4, these applications span mul-tiple domains such as workload distribution, cost-efficiency, green computing, and SLA compliance.

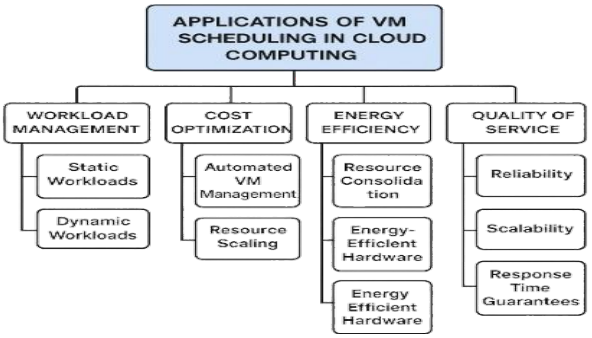

Fig. 4. Applications of VM scheduling in cloud computing

5 PROPOSED PERSONAL MODEL: INTELLIGENT VM SCHEDULING USING HYBRID HEURISTICS, MACHINE LEARNING, AND AUTOMATION

We initially present our suggested personal model that incorporates heuristic optimization, machine learning (ML), and DevOps-based automation into a common, intelligent scheduling framework. Our model is envisioned for the goal of intelligent, adaptive, and complete automation of VM scheduling in heterogeneous environments such as edge computing and fog environments, and multi-cloud platforms. The following figure 5 provides a descriptive overview of our proposed solution.

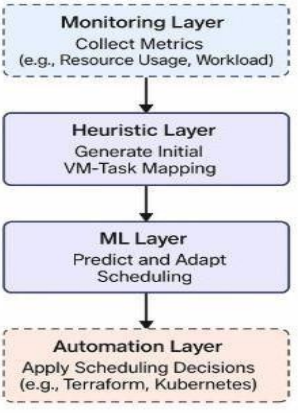

Fig. 5. Proposed hybrid model across four layers.

6 EXPERIMENTAL EVALUATION

To provide a preliminary insight into the potential effectiveness of our proposed hybrid scheduling model, we conducted a simple simulation comparing three virtual machine scheduling strategies: Round Robin, Least Loaded, and a Heuristic-Based approach. The simulation is lightweight and was implemented in plain Java without using simulation frameworks such as CloudSim, in order to quickly measure key performance indicators like makespan and average execution time.

The results, summarized in the table below, show that the Least Loaded method yields the lowest average execution time, while the Round Robin strategy provides a balanced makespan. The Heuristic-Based method, which introduces random assignment with basic load awareness, exhibits the highest makespan, highlighting its inefficiency in its current naive form.

Table 2. Comparative results of VM scheduling strategies based on Makespan and Average Execution Time.

Scheduling Strategy	Makespan (s)	Avg Execution Time (s)	Total Cloudlets
Round Robin	34	18.70	10
Least Loaded	34	18.50	10
Heuristic Method	48	20.80	10

These initial results serve as a proof of concept. A more comprehensive and large-scale evaluation using real-world workload traces and simulation environments such as CloudSim Plus or CloudAnalyst will be conducted in a follow-up article. That future work will also focus on implementing the complete version of our hybrid ML-enhanced scheduling model and benchmarking it against state-of-the-art intelligent scheduling algorithms.

7 CASE STUDIES OF VM SCHEDULING

Virtual Machine (VM) scheduling is a crucial aspect of cloud computing in order to attain high performance and efficient utilization of resources. Cloud service behemoths like Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure have devised various scheduling algorithms to schedule the workload.
In order to obtain a better view of the VM scheduling function in different cloud providers, a chart comparing the major cloud providers and their VM scheduling performance was provided.We will then go through real case studies of VM scheduling implementations, examining the scheduling methods used and the outcome obtained.

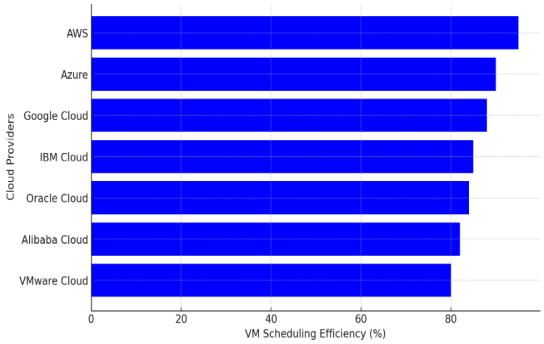

Fig. 6. VM scheduling efficiency across major cloud platforms

Major cloud providers have integrated intelligent VM scheduling to improve performance and scalability. AWS replaced its FIFO scheduling with Fair-Share Scheduling (FSS) in AWS Batch to balance resources among users . Microsoft Azure offers Spot VMs with cost savings, encouraging strategies like checkpointing to manage unexpected interruptions . Google Kubernetes Engine (GKE) uses priority-based scheduling and autoscaling to adjust resources dynamically . IBM Cloud applies AI-driven scheduling aligned with SLAs for predictive scal-ing , while Oracle Cloud uses Dynamic Resource Scheduling and preemptive scheduling to handle critical workloads . Alibaba Cloud combines ML models with its Elastic Compute Service for adaptive scaling . VMware Cloud integrates vSphere DRS and Proactive HA to migrate VMs in anticipation of hardware failures .

.

8 CHALLENGES IN VM SCHEDULING

VM scheduling in cloud environments faces multiple challenges, categorized as technical, practical, and emerging. Technically, scalability becomes difficult as cloud infrastructures grow, especially when managing heterogeneous tasks and resources . Multi-objective optimi-zation adds complexity by requiring trade-offs between cost, energy, and performance .

Practically, real-time adaptability is essential due to dynamic workloads and changing resource availability. Scheduling must also address security and privacy concerns, particularly with sensitive data .

Emerging challenges include the integration of edge and fog computing, which shifts resource management closer to the user and adds decentralization complexity . Additionally, fluctuating and unpredictable user demands require highly responsive and flexible scheduling mechanisms to maintain service quality ..

9 RECENT TRENDS AND FUTURE DIRECTIONS

Recent trends in VM scheduling focus on AI-driven decision-making, edge/fog integration, and sustainability. Future research explores multi-cloud optimization and energy-aware allocation.

9.1 Emerging Trends

Machine learning and artificial intelligence have enabled adaptive scheduling to provide the means of developing algorithms for real-time resource optimization, and hence system performance is enhanced [13]. Edge with wireless sensors under scenarios mobility and fog computing integration with VM scheduling are now essential because these paradigms enhance the traditional cloud infrastructures by bringing the resources closer to the sources of data, reducing latency, and improving efficiency [14][15].

9.2 Future Research Directions

Future VM scheduling must support multi-cloud and hybrid environments with better interoperability, pricing awareness, and failover support. Sustainability is also key, with growing focus on energy-efficient and thermal-aware scheduling to reduce the environmental impact of data centers.

10 CONCLUSION

VM scheduling is central to cost reduction, performance, and resource optimization in cloud computing. This paper reviewed major algorithm types—heuristic, metaheuristic, ML-based, rule-based, and hybrid—highlighting their trade-offs and suitability based on workload and infrastructure. Real-world use cases emphasize workload management, energy efficiency, and QoS. Cloud providers face challenges like scalability, real-time adaptation, and edge integration. In response, this paper proposes a hybrid model combining heuristics, machine learning, and automated DevOps to enable intelligent and adaptive scheduling. Future research will focus on self-optimizing, AI-driven systems capable of real-time, autonomous resource allocation.

References

1. 1. S. A. P. K. K. Supreeth, "Virtual Machine scheduling strategies in Cloud," International Journal on Emerging Technologies, pp. 181–188, 2019.

2. 2. P. Mell and T. Grance, "The NIST Definition of Cloud Computing," NIST Special Publication 800-145, 2011.

3. 3. S. Shivashankar, "Virtual Machine Scheduling Strategies in Cloud Computing – A Review," Reva University, 2019.

4. 4. D. A. S. N. J. A. Abdullah, "Load balancing techniques in cloud computing environment: a review," Journal of King Saud University – Computer and Information Sciences, vol. 34, no. 8, pp. 3910–3933, 2022.

5. 5. N. J. and N. Saied Asghari, "The Role of an Ant Colony Optimisation Algorithm in Solving the Major Issues of Cloud Computing," Journal of Experimental & Theoretical Artificial Intelligence, 2023.

6. 6. F. K. K. and S. Ghorashi, "Optimizing makespan and resource utilization in cloud computing environment via evolutionary scheduling approach," 2024.

7. 7. R. Yang, Y. Chen, T. Peng, and J. Xu, "Intelligent Resource Scheduling at Scale: A Machine Learning Perspective," 2018.

8. 8. "Priority based job scheduling technique that utilizes gaps to increase the efficiency of job distribution in cloud computing," Future Generation Computer Systems, 2023.

9. 9. A. D. Library, "Task Scheduling in Cloud Computing Using Hybrid Optimization," 2021.

10. 10. T. K. and K. Ajmera, "Energy-efficient virtual machine scheduling in IaaS cloud environment using energy-aware green-particle swarm optimization," International Journal of Information Technology, 2023.

11. 11. M. H. Sayadnavard, "A multi-objective approach for energy-efficient and reliable dynamic VM consolidation in cloud data centers," Engineering Science and Technology, 2022.

12. 12. H. Xu and X. Sun, "Fault tolerance and quality of service aware virtual machine scheduling algorithm in cloud data centers," The Journal of Supercomputing, 2022.

13. 13. G. Zhou, W. Tan, R. X. Sun, and S. Guangyao, "Deep reinforcement learning-based methods for resource scheduling in cloud computing: a review and future directions," 2024.

14. 14. Echoukairi, H., Kada, A., Bouragba, K., & Ouzzif, M. (2017). Evaluation and analysis of mobility scenarios with varying pause time for hierarchical protocols in wireless sensor networks. IOSR Journal of Computer Engineering (IOSR-JCE), 19(6), 62–70.

15. 15. Hmimz, Y., Chanyour, T., El Ghmary, M., & Malik, M. O. C. (2019). Energy efficient and devices priority aware computation offloading to a mobile edge computing server. In 2019 5th International Conference on Optimization and Applications (ICOA) (pp. 1-6). IEEE.