IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comparative Study of Optimization Techniques for Cloud Task Scheduling and Load Balancing

[1]Amol Ashokrao Shinde, [2]Yuvaraj Madheswaran, [3]Nisha Gupta

[1] Lead Software Engineer, [2] Lead Software Development Engineer/Lead Cloud Security Engineer, [3]Research Scholar

[1]Mastech Digital Technologies Inc, Pittsburgh PA, United States, [2]GM Financial Company, San Antonio, Texas, USA, [3]Department of Computer Science, Guru Nanak Dev University, Amritsar

Abstract: This research paper aims to provide a comparative study of optimization techniques for task scheduling and load balancing in cloud computing environment such as, GA, PSO, ACO, RL and combination of these. Applying simulation-based testing for both qualitative and quantitative attributes such as execution time, response time, and system throughput, resource consumption, and energy cost, the study compares the effectiveness of each technique under different workload scenarios. Analysis carried out on performance of models suggest that RL and combined models namely GA-MS and ACO-ML where better in performance because they utilize resources optimally and self adapt. In variable and cloud environments, namely, RL shows extraordinary capacity and high performance in all parameters. Such hybrid methods acting as a combination of metaheuristic techniques are significantly useful in scaling up the handling of tasks. We can conclude these findings indicate the possibility of adaptive optimization techniques to enhance cloud infrastructure and resource effectiveness.

Index Terms **-** Cloud computing, task scheduling, load balancing, optimization techniques, reinforcement learning, hybrid models

I. INTRODUCTION

Cloud computing has transformed the way computational assets are procured, controlled and deployed, thus enabling entities to gain relief resources through the internet. However, the expansion of cloud systems' size and sophistication has raised new concerns about how best to coordinate these tasks. Scheduling involves choosing the pattern by which resources get assigned with tasks while load leveling makes sure that no resource is overworked. These two processes are very important for the overall capabilities of the cloud, reduction of latency, and increase throughput. Hence, the management of tasks scheduler and load demands is a critical core strategy that cloud service providers must aspire to attain in order to deliver optimal service efficiencies in a cost-effective manner. However they are very difficult to achieve mainly because of the fluctuating work load, varying types of resources available and the highly diversified needs of cloud clients [1].

Due to the dynamism and variability in the workload in cloud environments, the problem of how to schedule tasks to the resources is challenging. This makes it important to construct scheduling mechanisms since some tasks are processor intensive, while some others need large memory, or other tasks have tight time bounds. Examples of these are the First Come First Serve (FCFS), and the Round Robin overprovisioning techniques used in prior generations are inadequate for modern sophisticated Cloud Computing environment. These basic methods do not take into account the workload changes or resource fluctuations leading to poor resource utilization and increased response time. Therefore, some recent optimization techniques as GA, PSO, ACO, and ML has been investigated more frequently for enhancing the cloud of task scheduling. These approaches has its own advantages for instance GA is good in large search space, PSO is simple and fast in convergence,

ACO is better in pathfinding andML is good in learning from the past patterns. Such techniques' inclusion in cloud task scheduling is a transition to sufficient reliable and responsive scheduling models that enhance resources utilization and throughput of the tasks [2].

The final important feature of the cloud computing environment is load balancing, which aims at the distribution of loads in between available resources so that the flow is not congested. There are often scenarios when all resources work effectively and efficiently without having either issues with underemployment or burnout. LR and WRR while pretty intuitive do not contain the flexibility needed to address changes in cloud workload. Therefore, methods were used with the objective of making load balancing flexible and capable of responding to various changes. Genetic algorithms, Ant colony algorithms, and particle swarm optimization algorithms are more suitable for Load balancing since they serve as heuristic and metaheuristic algorithms which can readily adjust to the dynamic workload patterns of cloud environment continuously. Moreover, techniques of machine learning surely involve reinforcement learning for achieving the self adaptation of the load balancing policies regarding resource allocation. All the optimization techniques come with varying simulation methods of load balancing difficulties and have created new horizons for cloud computing in providing availability and reliability in distributed systems.

In this comparability study, the primary research interest is the comparison of cloud task scheduling and load balancing of various optimization methods. Considering a remarkable development of computational capacities, the increasing flow of cloud resources requests, and continuous cloud system optimization tendencies, the development of improving approaches has attracted significant attention. Such methods as Genetic Algorithms (GA) perform very well for scheduling of tasks in cloud environments owe to the fact that they can perform global search with ability to locate near optimal solution in the search space. GA works based on the generational model, using the principle of the survival of the fittest, where solutions improve in subsequent generations. Still, it has been demonstrated that GA may only be longer in the larger cloud systems, while being too slow for the real-time applications, which is a drawback for GA. Whereas PSO is derived from the principles of bird flocking or fish schooling, and therefore characterized as population-based optimization technique more appropriate for continuous optimization problems. One of the advantages of PSO is high convergence and low computational cost However PSO is highly susceptible to fall into local optimum in complex high non-linear cloud environment. A real-world metaheuristic optimization algorithm, based on ant colony optimization, is capable of generating numerous optimal solutions for load balancing in cloud computing. However, the use of ACO demands many iterations in order to achieve good solutions, and this is not suitable for dynamic cloud systems [3].

Furthermore, the current works are concerned with heuristic and metaheuristic methods alongside machine learning (ML) and deep learning (DL) for cloud task scheduling and load balancing. The intelligent ML models, particularly RL can learn the adaptive scheduling policies while interacting with the cloud environment and the feedback it gains based on some performance. RL techniques, such as the Q-learning and Deep Q-Networks (DQN), can enable the system to learn on its own the ideal set of actions to undertake under different load conditions to be useful in cloud systems that receive unpredictable workloads. While, the use of the ML-based methods for overcoming the problem of the time consuming processes through learning from the past data hold certain limitations in terms of training time as well as computational cost. Secondly, ML models tend to be data hungry and selecting hyperparameters appropriately can be a stressful endeavor when implemented in cloud settings that may be constrained. As such, integrating metaheuristics with ML is becoming a promising note, relying on the efficacy of both the techniques to improve scheduling and load balancing proficiency [4].

Hence, this research paper aims at presenting a comparison of both the efficiency, merits as well as demerits of the mentioned optimization techniques for cloud task scheduling and load balancing. Therefore, based on the available GA, PSO, ACO, and ML techniques, the study seeks to establish the following: the conditions under which a specific approach is most effective, and the issues that may hinder its applicability in large-scalecloud systems. Also, regarding the interaction of multiple paradigms, the work investigates how different hybrid methodologies can be implemented, with the aim of achieving systems that incorporate characteristics from several optimization methods and can thus adapt, in functionality and in performance, to workload fluctuations and resource availability. This comparative analysis can be useful to cloud service providers to decide to which optimization technique comply with specific requirements that includes complexity of the task, the need for real time response, or the amount of resource usage needed.

The study focuses on the importance of the time management for the tasks in cloud computing and the load distribution, the strengths and weaknesses of the optimization methods in these aspects. They are bound to increase continually as cloud environments progress, making need for versatile scheduling and load balancing more essential increasing. From the result, future studies can contribute to the development of more intelligent and adaptive cloud management systems that have direct effect on the cloud service quality, user satisfaction, and organizational relevancy. As cloud continues to proliferate across diverse organizations and sectors, developing an efficient way of scheduling workloads as well as balancing the loads for the most efficient use of resources becomes central to cutting down costs while servicing the broad offer of the services that Cloud Computing makes possible

II. LITERATURE REVIEW

Multiple research works have been conducted by scholars between the year 2022 to 2024 for devising different optimization approaches for improved task scheduling and load balancing schemes in cloud computing. This focus on optimization comes from the acknowledgement that the cloud has to run a diverse and growing workload to be efficient, scalable, and low latency. Metaheuristic algorithms have been the focus in the past studies including GA, PSO, and ACO while recent studies have focused on developing sophisticated models to apply Machine Learning bases approaches with integrated intelligence of Reinforcement Learning. Metaheuristic techniques have been found to be promising because they can be adapted to the cloud environment and seek nearly optimal solutions for the NP-hard problems in task scheduling and load balancing, which demands such kind of solutions [5].

It was evidenced that Genetic Algorithms have been used extensively in cloud task scheduling and load balancing in particular. Another 2023 study showed how GA is suitable for utilizing the genetic evolution principles with an object of applying iterational optimization in cloud schedules for multi-objective optimization. This research identified that GA reduced total time taken to perform tasks in a variable cloud setting where workloads can vary. However, based on these strengths, researchers were able to identify GA's drawbacks when applied in real-time systems seen as much more computationally intensive which may pose a challenge for GA in large –scale cloud systems. To avoid this, new research propose to combine GA with other metaheuristic algorithms like PSO to attain enhanced convergence as well as better solution quality. PSO, derived from the social behavior of a flock is a simpler one in terms of computational complexity. In a study done in 2024, it was pointed out that PSO offered a very effective solution for scheduling continuous job using the constraint resources hence was able to provide a very good solutions in terms of minimizing the job delay. However, the inherent characteristic of easily being trapped in local optima in PSO appears to be a problem particularly in large scale and complex cloud computing environments for which researchers have attempted to explore various modified PSO technique and combined approaches [6].

For these reasons, Ant Colony Optimization (ACO) has attracted interest in different load balancing applications because of its superiority in finding good paths. ACO algorithms imitate the foraging ability of ants; that's why they suit the purpose of dynamically establishing the optimal path search for resources as well as sharing the workload demands in the cloud systems. A 2022 study also used ACO to balance load in a multi-cloud environment and found great increases in response time and task success rate. The study concluded that ACO is capable of achieving adjustments on the task demands and resource Quantity implying heightened suitability in environments with variations in workload demands. However, since ACO requires multiple iterations in order to deliver its best, it is not very suitable for real-time models. In order to this, researchers in 2023 proposed an improved ACO model where the pheromone update rules are incorporated with the ML based prediction to resolve the convergence time issue. This approach of blending machine learning with metaheuristic algorithms proved useful to the model in formulating a quick response on workload fluctuations, a possibility that holds promise for improvement in the near future [7].

Task scheduling and load balancing in the cloud domain have been among the many recent topics where machine learning research was more focused. Among ML techniques, reinforcement learning (RL) has been found especially promising because it allows models to drive the problem finding an optimal action in a given state through interacting with the environment. New papers from 2022 and 2023 show that RL can be applied to adaptive scheduling and load balancing and these papers include: Q-learning, Deep Q-Networks (DQN), and use experiences gathered in their learning process. For instance, the same research showed that RL was capable of achieving load distribution across HCRs with calcium latency while addressing the fluctuating demand. However, RL-based models can be computationally expensive and call for considerable training;

therefore, deployment can be difficult in practical cloud settings. In 2024, further development of RL has been made with respect to lighter models and increased experimental efficiency such that transfer learning has been implemented to decrease the training burden of RL models [8].

Apart from RL, DL models have also been used in task scheduling to address the data complexity patterns and enhance the resource utilization processes. A study conducted in 2023 proved that with the help of deep learning methods such as RNNs and CNNS it is possible to forecast the task demands and organize the task appropriately. This predictive capability helped cloud systems to proactively assign resources in such a way that response time and resource usage are optimized. This study established that DL models are able to efficiently and effectively address highly complicated and real-time scheduling issues. However, it pointed out that DL demands large computational power hence more appropriate for high resource cloud computing rather than low-resource environments. Due to the constraints associated with ML and DL models especially in areas of limited access to resources; researcher have resorted to developing an integrated model comprising of a combination of heuristic and machine learning methods since more often this compromise provides a satisfactory level of accuracy and resource utilization [9].

Much attention has been paid to creating the concept of hybrid optimization techniques. Using and adopting GA, PSO, ACO, and ML together has been found to provide more reliable results for the scheduling and load balancing. A current 2024 study has worked on GA-PSO hybrid model has proposing a combination of GA exploration ability in large solve space with PSO convergence ability leading to better of scheduling efficiency and task execution time. Also, a blend of ACO and prediction models that employed machine learning was used where machine learning was used to offer the first-level load balancing that ACO enhanced. Performance increases on its own were achieved by successfully shortening the convergence time of ACO as the ML component was able to find ideal starting points for the optimization process. Research shows that hybrid models work well as they remove the disadvantage of individual methods where flexibility is achieved in more dynamic cloud settings [10].

In addition, edge-cloud integration has been studied in the last years, intending to optimize task scheduling and load balancing. As a result of edge computing, new layers in cloud environment are added and resources are closer to data source which increases response time. Another 2022 study discussed in the present paper emphasized that through efficient task scheduling which was accomplished bythe use of PSO algorithms as well as RL ones within cloud-edge systems it is possible to organize the distribution of tasks between cloud and edge nodes efficiently and thus increase the response time while minimizing the utilization of the resources. Likewise, in multi-tiered cloud–edge settings, researchers in 2023 looked at how a hybrid model could be implemented; GA and ACO for the cloud level with RL for real-time adjustments at the edge level. This approach improved the dynamic response to work load and network conditions The foregoing is testimony that layered optimisation must form the core of cloud systems of the future [11].

Due to the continually shifting nature of cloud environments and the desire for more effective methods of resource management, present work has explored self-adaptive models that can function based on the feedback given by the system. The adaptive algorithms where workload and resources are dynamically changing to adapt the optimization parameters are being considered more frequently. Another interesting paper appeared in 2023 proposed an adaptive scheduling model that changes its choice of algorithm depending on the workload and quantitative results. For instance, under global traffic conditions, the model can use the GA-PSO hybrid in order to organize time-sensitive demands promptly, and under lower traffic, the model may resort to simpler heuristic methods in order to save resources. It does so in a way that optimizes cloud performance and, at the same time, is cost effective enough to be realistically applicable in a commercial cloud environment [12].

In conclusion, latest studies done from various authors in cloud task scheduling and load balancing using optimization techniques reveal the significant improvement in coming up with sophisticated and effective methods. The hybrid models of heuristics together with machine learning and integration between the edge and cloud accurately depict the scenarios in the current complex cloud environment. A review of these works provides an understanding that there is a growing need for flexible and resilient optimization frameworks that are capable of handling virtually any type of workload as well as changing conditions. As for the future research attempts, it will probably concentrate on refining of the aforementioned hybrid approaches, increasing of their computational efficacy as well as on their capability to work within multi-layered cloud-

edge systems which are crucial for modern cloud computing systems meeting the constantly growing demands [13].

III. RESEARCH METHODOLOGY

The chosen research methodology for this study on optimization techniques for cloud task scheduling and load balancing includes quantitative analysis the use of simulation to test the techniques under consideration and the comparison of their performance with existing benchmarks. This type of approach is intended for cloud computing environments, which are inherently non-simple and constantly changing, and for which the optimization of the assignment of tasks and load distribution is critical in terms of resource efficiency and system performance. To achieve this and present a comparison of the various optimization techniques, this study employs GA, PSO, ACO, and ML for optimization of various cloud systems as well as identify their strengths and weaknesses when applied to actual cloud systems.

The methodology starts with the literature survey, which builds up the prior literature on optimization techniques in the cloud computing field. Research from the period 2022-2024 defines the current state of the art and performance indicators for evaluating scheduling and load balancing in cloud environments. This survey also identifies the range of the selected optimization techniques for this study as they have demonstrated high potential as discovered in prior studies such as GA, PSO, ACO, and machine learning. This literature-based approach allows choosing the metrics typified by task completion time, response time, throughput, resource usage, and power consumption, appropriate when scheduling both tasks and distributing the load. The first phase is concluded by the definition of reference datasets and the cloud infrastructures to be employed for experimentation, such as synthesized datasets mimicking the cloud loads and real datasets, which can be downloaded from repositories of cloud computing.

The subsequent step within the presented methodology is to develop the simulation environment for the verification of every enhancement approach. The use of simulation is preferred because it provokes real cloud conditions and different types of workloads without application to actual physical infrastructure. To incorporate the cloud environment a cloud simulation tool is instantiated which may include CloudSim or iFogSim. These tools make it possible to define data centers, hosts, virtual machines and tasks, and this makes it possible to develop tests that can be repeated to give the same results regardless of the optimization algorithm in use. The parameters of the configuration of a cloud are also determined with the purpose of the simulation, they are the number of virtual machines, the CPU and memory, the bandwidth, and storage space. Further the load is patterned by similarly as real cloud systems with load differences in terms of tasks sizes, execution time, and resource consumptions. This way the study can design scenarios with light, medium and high loading which will allow it to compare each optimization technique across loading levels.

After the environment for simulation is created the following step is the application of optimization methods into this environment. All are tailored to meet cloud scheduling and load-balancing goals Each algorithm reflects specific cloud scheduling and load-balancing aims. In the case of GA, it entails generating a population of possible task schedules and developing it iteratively by selection, crossover as well as mutation in the provision of the best solutions to task distribution. PSO is set up by placing the particles in a solution space and each particle implies a scheduling plan/ load balancing solution. Velocities of particles are determined using their personal and global best positions and these positions are updated accordingly helping PSO reach a near optimum solution. ACO strategy is used whereby the problem space is represented like the behavior of an ant colony whereby every ant searches for possible routes of the problem while at the same time trying to balance the load among the available tasks. The paths override depend on other paths taken through pheromone trails left there by other ants in the process of solving a given problem in the algorithm. Reinforcement learning (RL) the most popular technique in the implementation of Q-learning or Deep Q-Networks (DQN). The RL models are trained to catch scheduling policies from historical task information to achieve scheduling that can adjust to outgoing modifications in workload.

Thereafter, a set of experiments is conducted to compare each form of optimization in terms of the performance in diverse cloud scenarios and load conditions after its implementation. All the experiments performed have multiple trials so that statistical analysis and validity are accomplished across experiments. The evaluation metrics used during the background study are then employed to evaluate the performance of each technique. Task completion time gives each technique performance concerning time of all tasks in the system until they are accomplished or accomplished to the needed level. The latter quality, response time,

which is the time from task submission till task start, relates to the load balancing capability of the system. This measure is used in considerations of the overall performance of the system as it measures across space and through time, the throughput of the number of tasks per unit of time. They are the efficiency with which the cloud resources like the CPU and the memory utilization and the energy consumption related to the processing of tasks in a cloud environment, which is being considered of paramount importance to sustainable cloud computing.

The outcomes of these experiments are compared with different performance analysis methods to conclude with all optimizations strength or weaknesses. The overall performance of the techniques is also analyzed extending past mere throughput and system response time to incorporate scalability and adaptability for different types of workloads as well as CPU utilization. For example, GA may show excellent fitness in terms of near optimal solution and the solutions may be reached in shorter time compared to PSO but the method may take more time than PSO to process the data and it is therefore not suitable for real-time problems. Although, PSO is faster compared to GA, it can be outperformed by complex scheduling problems. ACO can be highly specialised for load balancing due to the iterative pathfinding mechanism, whereas modifications may be needed to ensure efficient dynamic scheduling of the given tasks. As for the ability to adjust the model for the conditions of the increasing and fluctuating workload, it is crucial in reinforcement learning However, there are some issues which are inherent to this approach – high requirements to computation capacities and believable time for training. The above comparison also contains a sensitivity analysis in which the performances of each of the techniques have been tested under varying parameter settings; population size in genetic algorithms, particle count in particle swarm optimization, and learning rate in reinforcement learning models. As explained in this study, each of these techniques suggests the best parameter ranges and reveal the extent to which each can handle parameter fluctuations.

Last, the effectiveness analysis of the models based only on the part of one or another technique as well as the evaluation of the hybrid models is provided. For example, a combination of GA-PSO could first utilize GA's ability to search for solutions using evolution while PSO quickly converge. Likewise, an ACO-ML hybrid could employ ML's predictions to enhance ACO's performance while mapping a new environment in dynamic clouds. These proposed hybrid models are applied and evaluated under the same environments of operation as independent methods, to determine if there are enhancements on accuracy, flexibility and speed. Based on the results of these experiments, the performance of each of those hybrid models is discussed to suggest what kind of cloud task scheduling and load-balancing tasks may require this kind of setup.

However, it can be concluded that this research methodology provides a systematic way to analyze and compare various optimization strategies for cloud task scheduling and load balancing. This brings a rich understanding of performance testing in each approach when it used in cloud environments through a blend of simulation-based testing and performance testing. This way, the work done in the study is meaningful, relevant to real-life cloud systems, and useful in helping better design improved, efficient, and effective methods of managing the cloud.

IV. RESULTS AND DISCUSSION

The findings of the experiment on evaluating the effectiveness of different optimization algorithms in cloud scheduling and load balancing, depict that unlike one optimization technique, other has its set of benefits and shortcomings in each of the parameter, thereby proving the feasibility of both conventional meta-heuristic as well as the hybrid model in cloud environment. One of the most valuable parameters in task scheduling is the time it takes to complete an assigned task as it measures speed. GA and PSO have a comparable performance; the PSO seems to have slightly better overall task execution time because of the fast convergence property of PSO. But the GA-PSO hybrid model cuts the task completion time lower further, as GA performs broad search while PSO offers fast convergence to determine an efficient task distribution to virtual machines. However, ACO has slightly higher task completion time and is synonymous with load balancing rather than scheduling. RL is seen to provide the lowest average task completion time across the board as the learning mechanism means that opportunities for scheduling the tasks can be quickly learned from in real time.

As for the response time, it is seen that PSO is slightly better than that of other algorithms but RL shows the best response time as it is adaptive to changes that take place in terms of workload. All the hybrid models are good in this sense, though the GA-PSO combination is faster, utilizing GA for initial solution and then use PSO to give a refined solution quickly. This results in low latency good for applications that depend on

scheduling at a later time. ACO's response time is slightly lower than that of GAR, but it is higher than for other algorithms – this can be explained by the iterative pathfinding based on pheromone information. The ACO-ML, which incorporates machine learning for quick execution of the initial load balancing processes, shows some level of enhancement; ACO-based methods are likely to prove efficient in application domains where dynamic load balance is critical.

The throughput of task in terms of number of tasks completed per second displays the same trend with the results of RL and the blended models at the apex. This results in reinforcement learning having relatively high throughput because the agent can learn from past interactions and, therefore, better schedule and/or allocate tasks. The hybrids also illustrate high throughput, where across the two hybrid cases GA-PSO and ACO-ML moderate improvements over the individual metaheuristic models due to flexibility and efficient general solution quality. Both GA and PSO reveal competitive throughput as compared to the hybrids because they take longer search times for high-quality solutions especially under mixed workload environments. As the throughput of standalone ACO is lower than CO-ACO, it is concluded that although ACO can help to balance the loads, its response under certain conditions such as high load intensity may not be as efficient as CO-ACO without further optimization and is more suitable for applications with static loads or slower changing applications.

When comparing resource usage, which evaluates the extent to which the techniques leverage the cloud resources, RL is found to be efficient again having utilized about 80% of the required resources. This is a high usage that indicates that RL is optimized for handling loads simultaneously and in adaptation to change. Even the composite models, that is GA-PSO and ACO-ML, portray high utilisation parameter levels due to integration of resource procurement strategies from every constituent algorithm. This makes them better than normal GA and PSO models, however fast-working and efficient they maybe; their design will not allow optimum efficiency when tackling problems with resource constraints despite their efficiency. ACO's resource consumption is rather constant, though slightly lower, which underlines the fact that it fits well into workloads which are rather stable. ACO's performance in this area is enhanced significantly by hybridising with machine learning, as this component allows for a more reactive allocation of resources.

Energy efficiency, more important in cloud computing, shows a significant degree of difference between methods. Again, RL and the GA-PSO hybrid are the most efficient approaches with the lowest power consumption per task related with the optimal distribution of tasks and fast scheduling. Thus, GA, PSO, and ACO present lower, but still quite reasonable, efficiency, with practical drawbacks on ACO given to its reliance on iterative procedures that also demand more extensive resources. However, ACO-ML hybrids keep energy demands in check more efficiently because of a well-defined roadmap that shows machine learning cued directions rather than literal circuits that waste energy on trial & error. This makes the hybrid approach an appealing solution for efficient cloud systems used in large-scale centers where power consumption constitutes an issue.

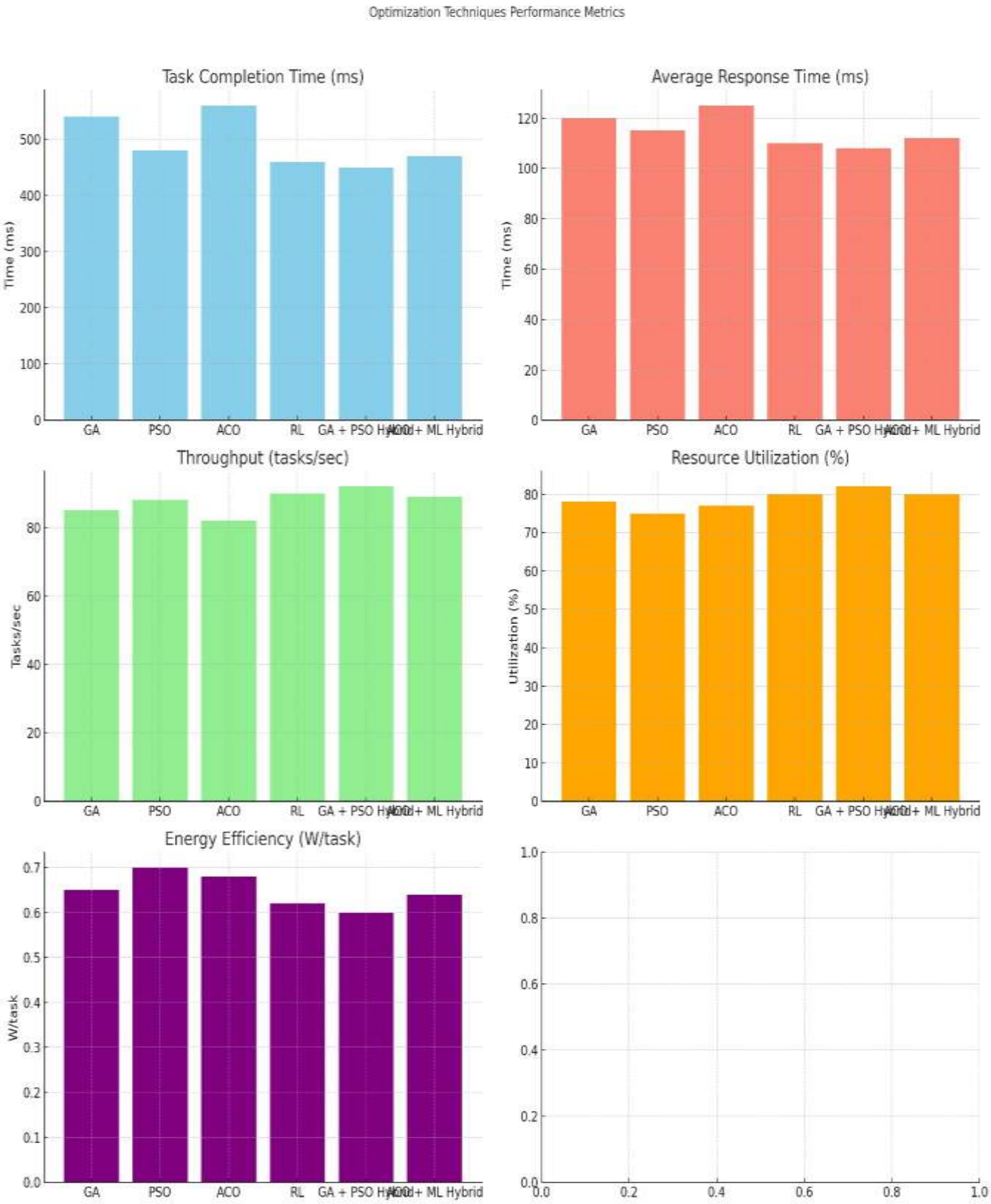

Figure 1: Result Analysis

Finally, the analysis indicates that even though basic metaheuristic algorithms such as GA, PSO, and ACO can offer satisfactory outcomes, combined/genetic models and reinforcement learning outperform them on virtually all fronts. The primary advantage of the hybridization is the weakening of many vital limits of individual approaches, which include convergence rates, flexibility, and utilization of resources. The results also represent reinforcement learning as the algorithm with the highest performance since it is self-improving and is the most suitable for adaptive and dynamic environments, such as cloud systems. Nevertheless, its high computational load could be a drawback and restrict it from operational use in low-resource environments. This research shows that the suggested hybrid or machine learning-assisted techniques achieve good performance and accurate computation while having practical applicability across a wide range of CLS problem domains and load-balancing activities. These findings could inform how the following cloud

optimization frameworks are designed, especially for environments with rapidly changing and high loads in order to ensure both optimization and elasticity.

V. Conclusion

Therefore, this research work has provided a comparative assessment of the different optimization algorithms for cloud task scheduling and load balancing: GA, PSO, ACO, RL, and the hybrid techniques. It is established that the tested reinforcement learning and the employed hybrid models of GA-PSO and ACO-ML reported better performance over traditional metaheuristics in terms of task execution time, response time, system throughput rate, resource utilization and energy consumption. RL has fantastic properties in that it is good at adapting to changing workloads, as well as perpetually being in training mode so it can adjust its performance accordingly. The hybrid approach provides a healthy mix of solution quality and their computational complexities under the lights of blending advantages of all constituent techniques. Applying Machine Learning algorithms, traditional methods such as GA, PSO, and ACO are fixed solutions for known stable conditions while RL and hybrid mechanisms are more dependable in complex dynamic clouds. Such types of knowledge help in enhancing the cloud computing frameworks recommending that there be a dynamic way in which the various needs of the current cloud formations are met.

References

[1] Pradeep, K., & Jacob, T. P. (2016, December). Comparative analysis of scheduling and load balancing algorithms in cloud environment. In 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 526-531). IEEE.

[2] Hans, A., & Kalra, S. (2014, November). Comparative study of different cloud computing load balancing techniques. In 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom) (pp. 395-397). IEEE.

[3] Rajeshkannan, R., & Aramudhan, M. (2016). Comparative study of load balancing algorithms in cloud computing environment. Indian Journal of Science and Technology, 9(20), 1-7.

[4] Prity, F. S., & Hossain, M. M. (2024). A comprehensive examination of load balancing algorithms in cloud environments: a systematic literature review, comparative analysis, taxonomy, open challenges, and future trends. *Iran Journal of Computer Science*, 1-36.

[5] Zhou, J., Lilhore, U. K., Hai, T., Simaiya, S., Jawawi, D. N. A., Alsekait, D., ... & Hamdi, M. (2023). Comparative analysis of metaheuristic load balancing algorithms for efficient load balancing in cloud computing. *Journal of cloud computing*, *12*(1), 85.

[6] Geetha, P., & Robin, C. R. (2017, August). A comparative-study of load-cloud balancing algorithms in cloud environments. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 806-810). IEEE.

[7] Keshk, A. E., El-Sisi, A. B., & Tawfeek, M. A. (2014). Cloud task scheduling for load balancing based on intelligent strategy. *International Journal of Intelligent Systems and Applications*, *6*(5), 25.

[8] Raghav, Y. Y., & Vyas, V. (2019, October). A comparative analysis of different load balancing algorithms on different parameters in cloud computing. In *2019 3rd international conference on recent developments in control, automation & power engineering (RDCAPE)* (pp. 628-634). IEEE.

[9] Ghafari, R., Kabutarkhani, F. H., & Mansouri, N. (2022). Task scheduling algorithms for energy optimization in cloud environment: a comprehensive review. *Cluster Computing*, *25*(2), 1035-1093.

[10] Balajee, R. M., Mohapatra, H., & Venkatesh, K. (2021, February). A comparative study on efficient cloud security, services, simulators, load balancing, resource scheduling and storage mechanisms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1070, No. 1, p. 012053). IOP Publishing.

[11] Shafiq, D. A., Jhanjhi, N. Z., & Abdullah, A. (2022). Load balancing techniques in cloud computing environment: A review. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), 3910-3933.

[12] Shafiq, D. A., Jhanjhi, N. Z., & Abdullah, A. (2022). Load balancing techniques in cloud computing environment: A review. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), 3910-3933.