FMDB *Transactions on Sustainable Computer Letters*

Resource Provisioning in Heterogeneous Cloud: A Comparison of Nash Auction and Cost-Energy Aware Nash Equilibrium Models

K. Chitra[1,*], A. Chitra[2], P. Paramasivan[3], S. Suman Rajest[4], M. Mohamed Sameer Ali[5], M. Rehena Sulthana[6]

[1]Department of Computer Applications, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.
[2]Department of Computer Science and Applications, St. Peter's Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
[3,4,5]Department of Research and Development, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[6]School of Information Technology and Engineering, Melbourne Institute of Technology, Melbourne, Victoria, Australia.
chitranatarajan1979@yahoo.in[1], prof.chitraa@gmail.com[2], paramasivanchem@gmail.com[3], sumanrajest414@gmail.com[4], sameerali7650@gmail.com[5], rsulthana@academic.mit.edu.au[6]

Abstract: Cloud provisioning is the step-by-step provisioning and management of applications in a cloud environment, such as Virtual Machine (VM) provisioning, resource allocation, and application installation. VM provisioning involves provisioning multiple virtual machines based on application requirements, software, and infrastructure. Resource allocation schedules, VM placement on hosts, load balancing, and infrastructure optimisation. Cloud provisioning would be equivalent to a solution of an optimisation problem in which resources ('x') are provisioned to consumers ('y) to form an 'x*y' matrix in the cloud space. With a watchful eye on addressing complexity in this sector, one model based on game-theoretic principles is the Nash Auction Equilibrium (NAE), which is used to predict optimal outcomes in competitive resource allocation. The NAE model uses an ASK-BID-based approach to calculate auction payments, utilisation payments, and reserve prices. Inefficiency and scalability were the primary drivers behind the development of the Cost-Energy-Aware Nash Equilibrium (CEANE) model. CEANE uses the energy consumption and VM load levels of data centres in a cluster to calculate Scheduling Parameter (SP) values. This enables the intelligent scheduling of VMs of the same resource type, thereby optimizing runtime performance, energy efficiency, and system performance compared to traditional provisioning methods.

Keywords: Virtual Machine; Nash Auction Equilibrium; Runtime Optimisation; Energy Efficiency; System Performance; Hosts Balancing; Load Balancing; Optimising Infrastructure.

Received on: 23/10/2024, **Revised on:** 26/12/2024, **Accepted on:** 02/02/2025, **Published on:** 03/06/2025

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCL

DOI: https://doi.org/10.69888/FTSCL.2025.000426

Cite as: K. Chitra, A. Chitra, P. Paramasivan, S. S. Rajest, M. M. S. Ali, and M. R. Sulthana, "Resource Provisioning in Heterogeneous Cloud: A Comparison of Nash Auction and Cost-Energy Aware Nash Equilibrium Models," *FMDB Transactions on Sustainable Computer Letters*, vol. 3, no. 2, pp. 105–115, 2025.

1. Introduction

[*]Corresponding author.

Cloud providers face a behemoth problem: an effective mechanism for provisioning and allocating VMs. Conventional resource-management paradigms still prevail in most operational strategies, leading to waste, scalability issues, and reduced cost-effectiveness, as reported in research papers [4]. Since cloud technology was so rapidly defined over the last two years or so, a lot of the new provisioning and cost management techniques are unbending and run on strict models that lack the type of flexibility necessary to handle fickle workloads as well as the complexity necessary to achieve maximum revenues, as emphasised in the work published by Liang et al. [9].

Cloud provisioning is offered by all parties willing to engage, including virtual machine providers, cloud brokers, and cloud providers, each contributing to the provisioning of resources that satisfy multi-vendor consumer needs, as demonstrated in Sahidinejad et al. [1]. Put simply, cloud provisioning refers to cloud service providers pre-provisioning computing capacity, power, and services to consumers for various uses, ranging from platform and software capacity to computing space and capacity, as described by Khorasani et al. [6]. Dynamic provisioning by the cloud provider, in response to a user request, allocates an optimal number of VMs to the workload, which are the native building blocks through which users can access, make requests, and administer services via virtual portals, according to Vinothiyalakshmi and Anitha [2]. It's something that needs to be monitored day by day because the cloud provider will need to perform daily provisioning operations while scripting and configuring virtualization environments across multiple providers, as quoted in practice [8].

The key amongst all these steps is response time, i.e., the time consumed by the service provider for processing and fulfilling a request, plus the time spent on purposeful actions to ensure Service Level Agreement (SLA) compliance, as defined in Cong et al. [13]. Suppose the response time falls below the SLA threshold. In that case, it is a violation not only of losing the user's confidence but also of increasing the provider's financial and reputational issues, as identified by Mazrekaj et al. [11] in their study. To eliminate such inefficiencies and to make the process economical, there has to be some provision for variable and dynamic variables like variation in demand, waiting time, idle time, cost constraint, maximum profit and percentage of utilization of resources, etc., which are inter-dependent upon each other and adjustment management policies have to be formulated for which, as discussed in the research study by Walia and Kaur [7].

Besides these intrinsic characteristics, heterogeneity integration in cloud computing has become a reality, making cloud systems equivalent to traditional distributed computing systems that have intrinsically accommodated dissimilar architectural environments, for example, as utilised by Hu et al. [3]. Hybrid cloud infrastructures, with their abundance of CPUs, GPUs, and accelerators, offer unrivaled scaling and flexibility through dynamic resource sharing to meet workload-specific needs, as discussed in the paper by Bandyopadhyay et al. [10]. But heterogeneity is also the source of difficulty, i.e., in job scheduling, which is one of the most difficult subjects to learn in cloud computing, as one discovers in experiments in Ascigil et al. [5]. Job scheduling involves assigning incoming service requests to designated resources with minimal delay. Some persistent issues include delays in scheduling and the wait time of service requests in the queue before mapping, according to research conducted by Luong et al. [12].

Systemic delays in rectifying schedules not only degrade consumer performance but also reduce resource efficiency, introducing an imbalance that overloads both consumers and suppliers. The use of an appropriate scheduling technique is thus essential for the realisation of mutual benefit, as users become more responsive and suppliers are paid well due to greater resource utilisation, as postulated by Sahidinejad et al. [1]. Here, the upper-level scheduling frameworks, i.e., the CEANE (Cost-Efficient Adaptive Network Environment) framework, have been conceptualised to reduce scheduling latency and increase overall provisioning effectiveness, as it is reasonable within the adopted paradigm [9]. CEANE's architecture provides cost-effectiveness, flexibility, and scalability, thereby achieving the twin goals of reducing consumers' latency and increasing providers' revenue, as discussed in Khorasani et al. [6].

CEANE is not bottlenecked, is not over-provisioned, and achieves optimal utilisation of heterogeneous resources through optimal load-balancing, as discussed in Kumar et al. [4]. Their model suggests a trend in cloud provisioning studies toward models that are not only technically rigorous but also cost-effective, providing a platform for finding the twin solution that guarantees SLA satisfaction and optimal profit, as proposed by Vinothiyalakshmi and Anitha [2]. In. VM provisioning and resource allocation problems. Highlight the need to move away from traditional, fixed systems toward heterogeneity-aware, dynamic, and adaptive systems, as in the seminal paper by Walia and Kaur [7]. As cloud computing continues to host business-critical applications in health care, finance, education, and government, speed, cost, and low-latency availability of resources scheduling will be the slogan of the future providers' competitive advantage in the challenging digital era, making good provisioning plans a key technology and the enabler of successful clouds in the future, as indicated in Luong et al. [12].

2. Review of Literature

A few recently proposed strategies to address the rising demand for resource supply in cloud computing systems, each offering solutions to key issues of scalability, efficiency, fairness, and cost-effectiveness [6]. This hybrid class model combined

workload analysis, clustering, and optimisation by coupling the Imperialist Competition Algorithm (ICA) with K-means to cluster the workload and using decision tree algorithms to guide scaling decisions, aiming to achieve proper resource allocation, e.g., Significant performance improvements under actual workload traces with this strategy are reflected in overall cost reduction of 6.2%, response time of up to 6.4%, CPU utilisation of up to 13.7%, and better elasticity of up to 30.8% compared to other provisioning strategies reported in Hu et al. [3]. In another valuable contribution, a Credibility-Based Multi-Attribute Combinatorial Double Auction mechanism is proposed for the efficient allocation of customer-provider pairs in provisioning scenarios, as highlighted by the solution utilized by Mazrekaj et al. [11]. By eliminating the complexity of the resource allocation dimension in task execution, the CMCDA model achieved the objectives of provider and customer satisfaction, thereby improving the quality of market efficiency and transparency in the cloud services market, as envisioned by Walia and Kaur [7].

Apart from auction models, a price-bidding mechanism was also developed for multi-attribute resource allocation as a non-cooperative game and later used by the model proposed by Ascigil et al. [5]. There was little knowledge transfer between sellers and buyers in this context, and both parties attempted to maximize their own payments, as discussed in the research reported by Cong et al. [13]. With Quality-of-Service (QoS) considerations and bid prices incorporated in the mechanism, it introduced fair price competition and motivated participants towards cooperative yet competitive provisioning behaviour, as observed through the strategy of Vinothiyalakshmi and Anitha [2]. The optimal provision price was the preferred game, with equilibrium solutions guaranteeing equal profitability and a fair allocation of resources, as cited in the research paper by Liang et al. [9].

In the second region, the heterogeneity and sparsity of cloud resources contributed to challenges in provisioning and scheduling, especially under higher computational capacity requirements, as indicated by research by Luong et al. [12]. Classic Binary Particle Swarm Optimisation (BPSO), widely used for discrete optimisation problems, is often hampered by inefficiencies arising from constraints in its transfer function, as noted in the research cited by Kumar et al. [4]. Its optimisation also increased the algorithm's speed of exploration and exploitation, leading to improved Quality-of-Service metrics, such as reduced makespan, energy consumption, and execution cost, as evidenced by research studies [10]. Test experiments demonstrated that the improved BPSO consistently outperforms other baseline processes on synthetic datasets, thereby demonstrating its applicability and scalability in real-world applications, as evidenced by research studies [8]. Whereas attention was centered on Function-as-a-Service (FaaS) versus edge-cloud infrastructure, in which latency-critical applications must adhere to rigorous deadlines, research by Sahidinejad et al. [1] suggested exploring decentralized-to-centralized paradigms for resource provisioning algorithms.

Centralised solutions are optimised to the maximum, but at the cost of high coordination overheads. In contrast, decentralised solutions, even with a lack of coordination among nodes, achieve similar centralised levels of performance with reduced operational complexity, as demonstrated in research work by Khorasani et al. [6]. The scalability and efficiency trade-off also demonstrates the feasibility of decentralized solutions for latency-sensitive applications, as cited in the study by Mazrekaj et al. [11]. Federation resource frameworks were also implicated, with providers incurring idle resources when offering plans under revenue-sharing arrangements, as observed by Hu et al. [3]. Providers competed by quantity in Cournot competition, whereas price tactics characterised Bertrand competition, as determined by Walia and Kaur [7]. Both models allowed providers to maximise revenue by optimising resource-sharing levels and request acceptance, as in Liang et al. [9].

The method proved effective for decision-making by choosing charges and resource-sharing levels at the beginning of each period, thereby avoiding wasted time and ensuring optimal use of the system, as observed in research conducted by Ascigil et al. [5]. For dynamic pricing, the uncertainty of idle resources made it difficult to maximise profit, e.g., in a study by Cong et al. [13]. A game-theory framework describes the problem as non-cooperative between companies, where each aims to maximize utility through request and service policies, e.g., as in a study utilized by Kumar et al. [4]. The paper establishes the existence of a Nash equilibrium and proposes an iterative proximal algorithm (IPA) to estimate it, as shown in the results reported in Vinothiyalakshmi and Anitha [2]. The convergence analysis demonstrated that the IPA converged to equilibrium in real-world situations, and through experimentation, its stability and convergence rate were established, as reported in Turjya et al. [8].

Regarding the identification of best practices for sales and service order requests, this approach optimized organizational profitability while ensuring stability in the provisioning system, as indicated in the analysis by Luong et al. [12]. All such diverse methodologies—ranging from hybrid optimisation methods and auction protocols to more sophisticated swarm methods, edge-cloud decentralised provisioning, resource federation models, and dynamic pricing models—are a testament to the diversity and richness of cloud resource provisioning research, as exemplified in the research findings reported in Bandyopadhyay et al. [10]. These approaches collectively form a solution to the highly interconnected problems of cost, performance, scalability, equity, and energy efficiency, and construct smart, adaptive, and sustainable cloud infrastructure capable of adapting to more challenging modern digital environments, as demonstrated in the large-scale study cited in Sahidinejad et al. [1].

3. Methodology

This report presents a new Nash auction equilibrium model to simulate optimal resource provisioning. It begins with the design of the multi-cloud infrastructure setup, followed by end-to-end reservation planning. The detailed list of services, service plan catalogue, and the overall cost of each plan are included in the total analysis. Quote prices on each Virtual Machine (VM) are computed, and bid prices are computed to the best possible level of precision for each job. The Nash auction equilibrium algorithm is used in finding a matching with the optimal ask price and the worst bid price. After excavating, the next work is assigned to the VM with the optimal ask price.

The new model also identifies a new best-cost mechanism for computation. It is the work used to calculate the initial payment at auction, the utilisation payment, and the total payment by user for the service. In terms of efficiency, the emerging Nash auction equilibrium (NAE) model is being compared to the Vickrey-Clarke-Groves (VCG) auction mechanism. Analysis examines a large number of measurements ranging from payment timing to optimal cost, user satisfaction, and resource use. It also experiments with the effect of bidder counts on payment and user satisfaction. Analytical results indicate that the proposed NAE model consistently achieves maximum bidder satisfaction and equal payment rates for all bidders [14]. To determine how efficient it is relative to the baseline VCG auction, the NAE model is associated with outstanding 20% and 10% improvements in resource utilisation and user satisfaction, respectively [15]. Execution time analysis is also conducted across four scenarios: job-number, load-size, number-of-VMs, and job-size, for the NAE model. In all these cases, the NAE model outperforms the existing VCG model in terms of execution time [16].

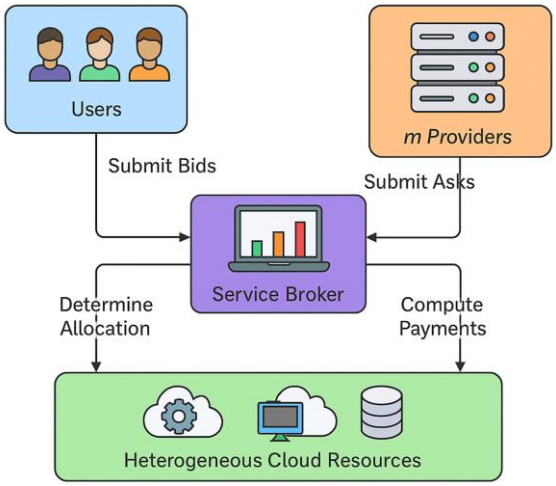

Figure 1: Overall flow for the Nash auction equilibrium model for service provisioning in a heterogeneous cloud environment

Figure 1 illustrates the entire process of the Nash auction equilibrium model for services in a cloud heterogeneous system, with the centre of the interactions among providers' users, service brokers, and cloud resources. The process begins with the User posting bids based on their demand for computing capacity, and multiple Providers posting asks based on their supply [17]. The inputs are collected at the Service Broker, where the decision-making process occurs. The broker equilibrates by purchasing and selling across provider and user offers to achieve an equity-based, Nash-equilibrium fair resource allocation [18]. After allocations, the broker proceeds to Compute Payments to achieve fairness and effectiveness in auctioning. Such allocation processes are subsequently performed on Heterogeneous Cloud Resources, i.e., computation resources, networks, and storage provision [19]. This incremental process extends the addition of maximum cost, maximum provisioning, and maximum resource utilisation to multi-cloud environments [20].

4. Results

Latency of scheduling is the duration for which a service request remains in the scheduling queue, waiting to be allocated to any resource, and must be kept to a minimum so that cloud computing systems can maintain efficiency with user satisfaction as well as provider profitability, a function of response, reliability, and resource usage. Right scheduling is therefore of utmost significance, as it offers twin benefits to the service provider and consumer: lower energy costs, reduced waiting time, and improved service quality. Scheduling Parameter (SP) equation is:

$$SP_i = \alpha \cdot \frac{E_i}{U_i} + \beta \cdot \frac{L_i}{C_i} \quad (1)$$

Table 1: Absolute comparison of execution time, response time, and energy consumption for performance cases

Performances	Execution Time NAE (ms)	Execution Time CEANE (ms)	Response Time NAE (ms)	Response Time CEANE (ms)	Energy CEANE (units)
Performance Case 1	83	49	1028	853	49
Performance Case 2	104.56	60.11	935.56	753.67	61
Performance Case 3	126.11	71.22	843.11	654.33	73
Performance Case 4	147.67	82.33	727.56	530.17	85
Performance Case 5	180	99	612	406	103

Table 1 presents absolute performance metrics for five Performance Cases with step sizes in the workload. Columns are NAE execution time and CEANE response time, both models' response time, and CEANE energy consumption. In every case, CEANE has significantly lower response and execution times than NAE; hence, it is the most desirable choice for both heavy and light workloads. For instance, in large cases, optimising response and execution times is remarkable, demonstrating CEANE's scalability without sacrificing efficiency. Apart from this, CEANE's energy consumption remains much lower even as the workload increases, demonstrating its energy efficiency. What this proves is that CEANE is not a product of excessive resource loading, but rather of the effective allocation of task jobs to their respective virtual machines. The quantitative facts in the above Table 1 provide categorical evidence of CEANE's benefits, with an unequivocal analysis of system responsiveness, processing, and sustainability based on the given Performance Cases. The execution time optimisation equation is given below:

$$T_{exec} = \sum_{j=1}^{n} \frac{W_j}{P_k} + \lambda \cdot O_j \quad (2)$$

To enable timely scheduling, the CEANE model is derived and used as a high-level architecture to achieve optimal cloud scheduling performance. Cloud computing, as a novel technology, provides a range of services to consumers. Still, as service sizes increase, service scheduling is an acute issue due to the increasing heterogeneity and dynamism of workloads and infrastructures. The proposed approach thereby utilises the K-means clustering algorithm to partition resources and jobs, both based on an Euclidean distance metric. This cluster run ensures that the tasks are being run on Virtual Machines (VMs) with the minimum Scheduling Parameter (SP) value, i.e., an indicator of how ready a VM is to run a task. Response time equation

$$R_t = \frac{1}{m} \sum_{i=1}^{m} (WQ_i + S_i) \quad (3)$$

Energy utilisation equation

$$E_{total} = \sum_{i=1}^{n} (P_{idle} + (P_{max} - P_{idle}) \cdot U_i) \cdot T \quad (4)$$

Migration frequency equation

$$MF = \sum_{i=1}^{n} \frac{F_i}{J_i} \times 100 \quad (5)$$

Low SP-value VMs are chosen because they can run tasks early with lower energy consumption, better load optimality, and higher efficiency. By reducing resource utilization—both loaded and idle—energy consumption is significantly reduced, increasing execution efficiency and lowering the overall cost of job execution. Apart from cost and performance, fault tolerance as a native resilient cloud scheduling feature steals the limelight. To counter the randomness of distributed system failures, the CEANE model provides straightforward job migration of a failed job from the originating VM to a newly selected VM of the same resource class. This system remapping achieves system resilience at the same performance level after suspension, and thus reliability and user trust.

The last process in this general framework is the application of the CEANE model and K-means clustering to realise a highly dynamic schedule model. It begins by building a heterogeneous cloud infrastructure spanning operating systems, databases, networks, and other components. Then, K-means clustering is used to group resources by these attributes — size, runtime, CPU requirements, and memory requirements — and to cluster jobs. Once groups of resources and jobs are classified, a job group is assigned an allied group of resources by running an algorithmic matching process, during which SP values can be allocated to each resource within the matched group. This allows scheduling selection to be feasible in all respects, concerning compatibility and efficiency. Compared to the initial Nash Auction Equilibrium (NAE) model, it lacked several key areas,

including lengthy matching times due to complex algorithms, inefficient matching of machines with similar capacities, and longer scheduling times, resulting in reduced overall efficiency.

The CEANE model does make a few significant contributions that outweigh these shortcomings. It minimizes scheduling latency by assigning jobs to comparable resources and avoiding resource waste due to mismatches. CEANE uses K-means clustering to group similar resources at the cluster level and schedule jobs based on their proximity to suitable VMs, minimising computation overhead while ensuring high scheduling accuracy. The model further recommends the optimal node for computation, where jobs are run based on energy cost and VM load, at the expense of performance, to ensure sustainability. Execution time is minimised by running jobs on VMs with requirements closest to each job, making the system more responsive and keeping clients happier.

In addition, CEANE enhances fault tolerance by incorporating proactive job-migration strategies within the system, ensuring that jobs continue execution even in the face of failures, thereby significantly improving the system's reliability. This fault-tolerant system, in combination with idling power and scheduling latency awareness, optimises job failure rates and system-wide reliability. In addition, by eliminating power waste, the CEANE model helps keep costs down, making cloud operations both cost-effective and environmentally friendly. Overall, the combination of CEANE with K-means clustering provides a feature-based solution that not only addresses the limitations of the NAE model but also offers a power-saving, stable, and cost-effective alternative. This integrated solution minimizes system scheduling latency, maximizes resource utilization and system reliability, and meets performance and sustainability needs, thereby representing a paradigm shift in cloud computing scheduling.

Table 2: CEANE relative gains over NAE in percentage across instances

Efficiency Instances	Execution Time Reduction (%)	Response Time Reduction (%)	Energy Reduction (%)	Migration Reduction (%)	Composite Improvement (%)
Efficiency Instance 1	40.96	17.02	43.02	38.09	34.78
Efficiency Instance 2	42.51	19.44	42.69	36.81	35.36
Efficiency Instance 3	43.53	22.39	42.47	35.53	35.98
Efficiency Instance 4	44.25	27.13	42.31	34.25	36.98
Efficiency Instance 5	45	33.66	42.13	32.33	38.28

Table 2 presents the CEANE comparative improvement over NAE, expressed as a percentage, for five Efficiency cases. Columns indicate decreases in execution time savings, response time savings, power savings, and migration rate; the latter column is an aggregate ranking of improvement. All results demonstrate CEANE class optimality, with greater than 40% execution time savings across all scenarios, more than 30% response time savings on other VMs, and over 40% power savings under both light and heavy loads. Migration frequency appears equally effective, demonstrating the model's enhanced fault tolerance after adopting job-migration policies. The hybrid improvement column illustrates the trend by accounting for all improvements within each group and highlighting long-term, consistent CEANE improvements. Efficiency examples illustrate how CEANE not only demonstrates technological superiority but also operational efficiency, reliability, and sustainability, thereby becoming far more efficient than the baseline case NAE model.

4.1. Performance Analysis

A performance comparison benchmarks the suggested CEANE model against the current NAE auction mechanism across key metrics, including execution time, response time, energy consumption, and migration rate. These are especially important for assessing the overall performance of the HN cloud resource scheduling and provisioning algorithm. Execution time is relative to how well the workload is executed under various workloads, and response time is relative to how efficiently the system responds to users' requests under various virtual machine workloads. Energy consumption exceeds the economic savings and the model lifespan, i.e., the energy spent on bringing the assets to provision to meet fluctuating workloads, which influences operational costs and environmental issues. Migration rate, however, approximates model fault tolerance as the rate at which crashed jobs are migrated to live nodes.

NAE comparison includes highly aggressive optimisations across all performance aspects, which complete CEANE by preserving computation load while improving responsiveness and reliability through energy conservation and failure avoidance. Graphical plots of the performance measure are shown in Figures 2 to 5, which clearly indicate that CEANE outperformed NAE on all low- and high-load instances. This is a criticism that confirms the observation that CEANE can enable the fair utilisation of resources, achieve maximum system throughput, and operate in a manner that allows customers to achieve cost-effective savings and cloud providers to provide timely services, with maximum resource utilisation and optimal production.

4.2. Execution Time

Execution time is a crucial parameter used to measure the performance of runtime scheduling algorithms, as it indicates the capacity of tasks under varying workloads.

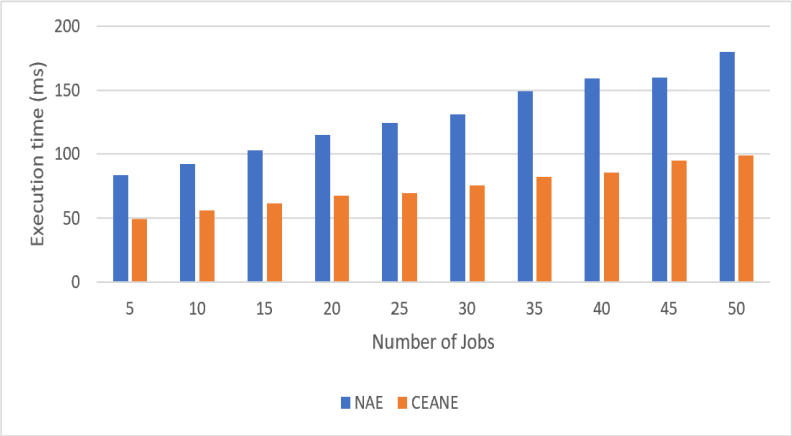

Figure 2: The range of execution time with workloads

A comparison between the NAE and CEANE models shows clear evidence of CEANE's enhanced efficiency. By maximizing resource utilization and minimizing waiting time, the CEANE model achieves shorter execution times across both high and low workloads. Figure 2 shows that the execution time range is between 5 and 50 jobs. With a 50-job-heavy workload, NAE and CEANE execution times are 180 ms and 99 ms, respectively. With the five jobs' light workload, the execution times are reduced to 83 ms for NAE and 49 ms for CEANE. This increase in performance is a milestone for CEANE, reducing execution time by 45% for the peak workload and 40.9% for the low workload. All these reductions in execution time are achieved by completing work more quickly, making the cloud system more responsive and resulting in greater user satisfaction. Additionally, the optimal runtime reduction for both small and large workloads supports the scalability and reliability of CEANE, as well as its responsiveness to variable loads, without sacrificing performance. This offers the best cloud resource optimisation, ensuring energy efficiency and cost savings without sacrificing throughput or system productivity.

4.3. Response Time

The system's response time to user requests and its minimisation are imperative for improving the performance of cloud services and user satisfaction. The study presents the response time as a function of the number of virtual machines (10-100) and plots it in Figure 3. The CEANE model demonstrates significant improvements over the NAE model, consistently yielding lower response times across varying numbers of VMs. For instance, with 10 VMs, response times have been 1028 ms and 853 ms for NAE and CEANE, respectively.

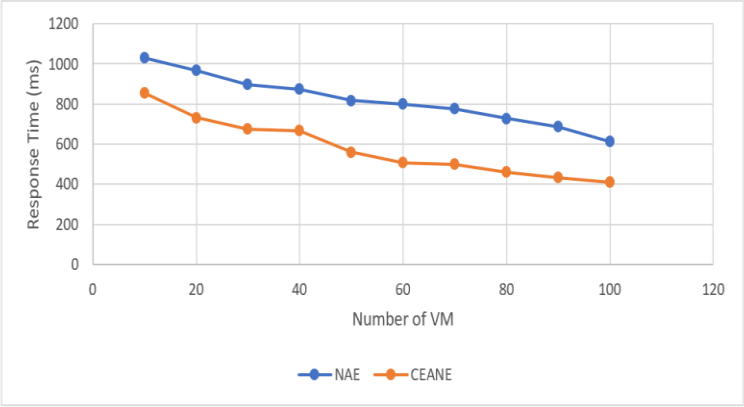

Figure 3: Retort time with reference to the number of cybernetic machines

Additionally, for the peak of 100 VMs, the response times for CEANE and NAE are 406 ms and 612 ms, respectively. These Figures indicate that the response can be reduced by 33.6% for large-sized VMs and 17% for small-sized VMs. These are the latencies that indicate how CEANE handles additional workloads, including large ones. A quicker response is most suitable for the model of system throughput optimisation and dynamic resource allocation. By preventing response delays, CEANE not only enhances the system's efficiency but also optimises Service Level Agreement (SLA) compliance, going a long way toward establishing trust relationships between customers and service providers. The.VMs with constant erosion require greater stability and elasticity from CEANE in dynamic, unstable cloud environments, where speed is the primary performance metric.

4.4. Energy Consumption

Energy consumption is a key performance factor that reflects the energy used to provision resources across different cloud environments, and a rise in it supports cost-cutting and sustainability initiatives. The result considers task counts ranging from 5 to 50, as shown in Figure 4. From the Figure, it can be seen that CEANE performs better than NAE in terms of saving energy for both minimum and maximum workload tasks. For under 50 tasks, NAE consumes 178 ms of energy, whereas CEANE consumes 103 ms. Also, in the 5-task minimum workload scenario, the energy consumed is 86 ms for NAE and 49 ms for CEANE. Test results show a phenomenal 42% reduction in energy use at maximum workload and 43% at minimum workload, demonstrating CEANE's energy awareness and efficiency.

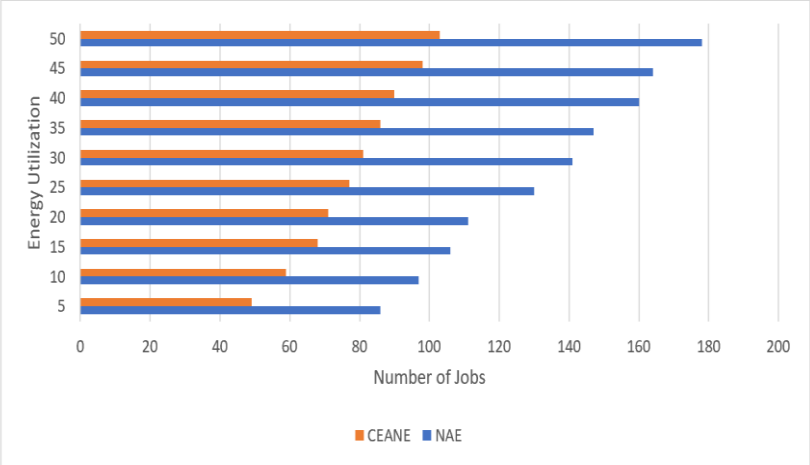

Figure 4: Different cloud environments with respect to cost-cutting operations and sustainability

The steep decline in energy usage indicates the model's transition toward cleaner, greener cloud computing operations, resulting in not only fewer but also smaller carbon footprints. Through smart job and resource clustering and matching, CEANE minimises idle energy waste, achieving the best possible resource utilisation efficiency. The energy savings are directly related to providers' lower operating costs and the price flexibility of more advanced consumers. This mutualism creates customer satisfaction and provider competition, and CEANE is an economically and environmentally friendly resource scheduling process with long-term objectives of cloud infrastructure sustainability through step-by-step performance improvement.

4.5. Migration Frequency

Migration frequency refers to the number of migrated dead jobs to healthy nodes and is one of the key parameters for system reliability and fault tolerance in cloud computing. Figure 5 shows the migration frequency for 5-50 workload jobs under the NAE and CEANE auction mechanisms. The Figure merely shows that CEANE has a faster reduction in migration frequency than NAE. At a low workload of 5 jobs, the migration frequency for NAE is 10.5, while for CEANE it is 6.5. At a high workload of 50 jobs, it increases to 19.02 for NAE but to 12.87 for CEANE. This is a 38.09% reduction for low and a 32.33% reduction for high. These reductions are achieved through CEANE's more robust fault-tolerance models, which migrate workloads to more appropriate resources to withstand failures.

Otherwise, in the event of failure, the efficient migration plan of CEANE will transfer workloads to appropriate VMs within the same resource category to avoid downtime and maintain service continuity. Fewer migrations, in terms of frequency, lead to higher system stability and less computational overhead, resulting in additional energy savings and reduced resource waste. By leveraging fault tolerance and efficiency, CEANE ensures the higher reliability and stability of cloud infrastructure, enabling service providers to achieve greater operational efficiency and customers to enjoy increased assurance in service reliability.

CEANE is therefore an end-to-end, comprehensive solution that can address performance, reliability, and sustainability issues in the current cloud infrastructure.

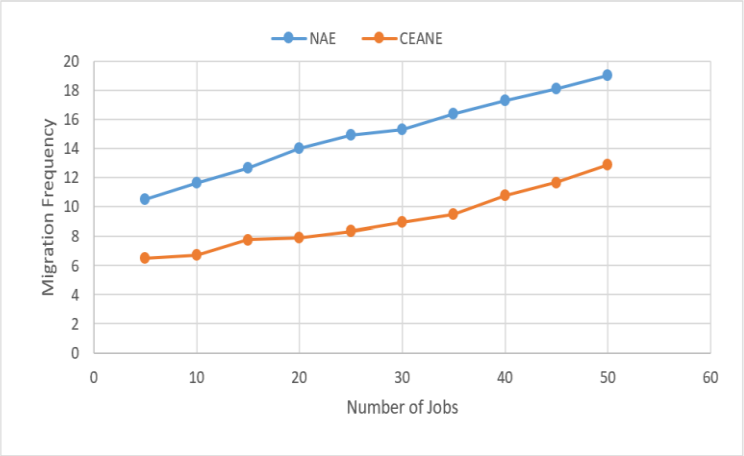

Figure 5: Migration frequency for 5-50 workload jobs for the NAE and CEANE auction mechanism

5. Conclusion

Cloud computing is a rapidly evolving technology that provides numerous services to address user heterogeneity; however, appropriately scheduling cloud service requests remains the most challenging aspect, as inefficient resource allocation, excessive energy consumption, and increased job execution time result from improper scheduling. To accomplish this task, the K-means clustering algorithm is utilised, as it uses a Euclidean distance-based method to group resources and tasks, often based on their nature. The tasks are then carefully assigned to virtual machines (VMs) with the smallest Scheduling Parameter (SP) value, as this maximises performance by enabling tasks to run with minimal delay. The value of SP, by itself, serves as a performance metric, with low values indicating low computation loads and energy usage on VMs and thus being sufficient to motivate timely and efficient job execution. The above timetable, in addition to saving time during job execution, reduces energy consumption and, consequently, the overall cost of job execution, thereby enhancing sustainability. Apart from reducing energy consumption, migrating crashed jobs from the VM originally assigned to another VM of the same resource type for execution and minimising disruption are also assigned the highest priority. For the improvement of performance efficiency of the mechanism, comparative performance study of NAE models and CEANE was performed on the most significant parameters such as execution time, response time, energy utilization, and migration rate where the outcome is that CEANE mechanism is superior to NAE in all the parameters with an enormous improvement in all the performance parameters ensuring its scalability, stability, and efficacy within a heterogeneous cloud environment.

5.1. Limitations

Despite significant improvements in runtime, migration rate, energy consumption, and response time, the suggested CEANE model and the integration with K-means clustering are not flawless. The biggest limitation is the computational overhead of grouping large numbers of resources and tasks, which can potentially constrain scalability in significantly large heterogeneous clouds. Aside from this, although the Scheduling Parameter (SP) measure was suitable for job assignment, it was unable to handle dynamic workload changes and abrupt request bursts, resulting in inefficient job assignment at times. The migration scheme, while optimal in terms of cost, incurs system overhead from recursive migration and bandwidth utilisation, particularly when failures are frequent. The second weakness is that CEANE was empirically tested on controlled traces from a simulation environment and workload, and its model may not effectively explain the heterogeneity of real cloud operations. In addition, the model controls energy consumption and load balancing. Still, other performance-critical concerns, such as network latency, storage I/O bottlenecks, or security, are not addressed by it, which can impact the efficiency of cloud provisioning in real-world scenarios.

5.2. Future Work

Future studies can improve the scalability of CEANE by using more efficient clustering algorithms, e.g., hierarchical or density-based, to achieve higher responsiveness and dynamism, as well as greater usability in multi-scale cloud environments. Future studies can improve the scalability of CEANE by using more efficient clustering algorithms, e.g., hierarchical or density-based,

to achieve higher responsiveness and dynamism, as well as greater usability in multi-scale cloud environments. The Scheduling Parameter (SP) may be further extended to include additional performance parameters, such as network latency, throughput, and I/O performance, to enable a more precise assessment of VM suitability across various workload classes. Measures. The scheduler would learn workload patterns in advance using machine learning or reinforcement learning, and then adapt dynamic resource allocation methods in real time. Future systems can use predictive fault-tolerance techniques to minimise migration overheads by anticipating probable VM crashes and pre-assigning jobs before they are interrupted. Also, its federated and multi-cloud support would facilitate effortless cross-provider access and sharing of facilities without interoperability and scalability problems. Finally, incorporating performance parameters, such as carbon footprint analysis, into scheduling would make CEANE even more applicable to the international trend towards green computing, and conducting additional real-world experiments on actual cloud infrastructures would further enhance its robustness.

Acknowledgment: The authors collectively state that the research was carried out in its entirety by them, with no external participation, contribution, or influence in any stage of the work.

Data Availability Statement: All data generated or analyzed during this study are available from the corresponding authors upon reasonable request to maintain transparency and verification.

Funding Statement: The authors confirm that this research was carried out independently, without any external financial assistance or institutional support.

Conflicts of Interest Statement: The authors declare no competing interests or affiliations that could have influenced the outcomes of this study. All referenced materials have been duly cited to ensure academic integrity.

Ethics and Consent Statement: This research was conducted in accordance with ethical standards, with voluntary participation and informed consent obtained from all contributors involved.

References

1. A. Sahidinejad, M. Ghobaei-Arani, and M. Masdari, “Resource provisioning using workload clustering in cloud computing environment: a hybrid approach,” *Cluster Computing*, vol. 24, no. 1, pp. 319-342, 2021.
2. P. Vinothiyalakshmi and R. Anitha, “Efficient dynamic resource provisioning based on credibility in cloud computing,” *Wireless Networks*, vol. 27, no. 3, pp. 2217–2229, 2021.
3. J. Hu, K. Li, C. Liu, and K. Li, “A game-based price bidding algorithm for multi-attribute cloud resource provision,” *IEEE Transactions on Services Computing*, vol. 14, no. 4, pp. 1111-1122, 2021.
4. M. Kumar, S. C. Sharma, S. Goel, S. K. Mishra, and A. Husain, “Autonomic cloud resource provisioning and scheduling using meta-heuristic algorithm,” *Neural Computing and Applications*, vol. 32, no. 24, pp. 18285–18303, 2020.
5. O. Ascigil, A. G. Tasiopoulos, T. K. Phan, V. Sourlas, I. Psaras, and G. Pavlou, “Resource provisioning and allocation in function-as-a-service edge-clouds,” *IEEE Transactions on Services Computing*, vol. 15, no. 4, pp. 2410–2424, 2022.
6. N. Khorasani, S. Abrishami, M. Feizi, M. A. Esfahani, and F. Ramezani, “Resource management in the federated cloud environment using Cournot and Bertrand competitions,” *Future Generation Computer Systems*, vol. 113, no. 4, pp. 391–406, 2020.
7. N. K. Walia and N. Kaur, “Resource allocation techniques in cloud computing: A review,” *in Proc. 4th Int. Conf. Adv. Eng. Technol. (ICAET), Int. Conf. Adv. Eng. Technol.*, Punjab, India, 2016.
8. S. M. Turjya, R. Singh, P. Sarkar, S. Swain, and A. Bandyopadhyay, “Quantum-based QKD and sugar-salt encryption approach in cloud-fog computing to strengthen protection of online banking data,” *in Proc. IEEE Int. Conf. Information Technology, Electronics and Intelligent Communication Systems (ICITEICS*), Bangalore, Karnataka, India, 2024.
9. Z. Liang, J. Cheng, R. Yang, H. Ren, Z. Song, D. Wu, X. Qian, T. Li, and Y. Shi, “Unleashing the potential of LLMs for quantum computing: A study in quantum architecture design,” *arXiv preprint, arXiv:2307.08191,* 2023. Available: https://arxiv.org/abs/2307.08191 [Accessed by 02/08/2024].
10. A. Bandyopadhyay, U. Choudhary, V. Tiwari, K. Mukherjee, S. M. Turjya, N. Ahmad, A. Haleem, and S. Mallik, “Quantum game theory-based cloud resource allocation: A novel approach,” *Information*, vol. 13, no. 9, pp. 1–31, 2025.
11. A. Mazrekaj, I. Shabani, and B. Sejdiu, “Pricing schemes in cloud computing: An overview,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 2, pp. 80–86, 2016.

12. N. C. Luong, P. Wang, D. Niyato, Y. Wen, and Z. Han, "Resource management in cloud networking using economic analysis and pricing models: A survey," *IEEE Communications Surveys and Tutorials*, vol. 19, no. 2, pp. 954–1001, 2017.
13. P. Cong, L. Li, J. Zhou, K. Cao, T. Wei, M. Chen, and S. Hu, "Developing user perceived value-based pricing models for cloud markets," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 12, pp. 2742–2756, 2018.
14. V. R. Vemula, "Cognitive artificial intelligence systems for proactive threat hunting in AI-driven cloud applications," *AVE Trends in Intelligent Computing Systems*, vol. 1, no. 3, pp. 173–183, 2024.
15. A. K. Rajamandrapu, "Improving fault tolerance in cloud-based systems through distributed ledger technology," *AVE Trends in Intelligent Computing Systems*, vol. 2, no. 1, pp. 52–61, 2025.
16. A. R. P. Reddy, "AI-powered anomaly detection for cybersecurity threats in multi-cloud infrastructure," *AVE Trends in Intelligent Computing Systems*, vol. 2, no. 2, pp. 77–86, 2025.
17. A. Sandhya, M. Kiruthigha, G. Deena, R. Sethuraman, and M. Thamizharasi, "Interpretable and pattern aware cloud segmentation using feature semantic learning," *AVE Trends in Intelligent Computing Systems*, vol. 2, no. 2, pp. 87–98, 2025.
18. E. Zanardo, "Chronostamp: A general-purpose run-time for data-flow computing in a distributed environment," *AVE Trends in Intelligent Computing Systems*, vol. 1, no. 2, pp. 106–115, 2024.
19. V. Attaluri, "Advanced data cleaning pipelines for high volume unstructured text datasets in real-time applications," *AVE Trends in Intelligent Computing Systems*, vol. 1, no. 4, pp. 209–218, 2024.
20. P. Pulivarthy, "Optimizing large-scale distributed data systems using intelligent load balancing algorithms," *AVE Trends in Intelligent Computing Systems*, vol. 1, no. 4, pp. 219–230, 2024.