FMDB *Transactions on Sustainable Computing Systems*

Optimizing Resource Allocation in Multi-Cloud Environments for Cost Efficiency and Scalability

Srinivasa Gopi Kumar Peddireddy[1,*]

[1]Department of Network Implementations and Operations, Charter Communications, Hutto, Texas, United States of America.
srinivasagopikumar.peddireddy@charter.com[1]

Abstract: Single-vendor-cloud minimization is the core to cost-performance equilibrium. Resource sharing among various providers is the aim of this research. With elastic software designs, predictive costing models, and dynamic load balancing, the method ensures optimum cost up to the point it is aligned with performance. Its methodology entails continuously monitoring resources, auto-scaling, and cost estimation tools. The experiment used Amazon Web Services (AWS) and Google Cloud Platform (GCP) data sets and built variables like instance cost, compute utilization rates, and response time. The proposed model uses Kubernetes and Terraform tools for automation to attain scalability in a non-intrusive manner and improve performance management. The experimental setup ensures the system's performance, enhanced cost-effectiveness, and scalability. The results confirm that strategic resource planning can make a 30% cost improvement and 25% performance. The study offers a critical reference model for companies seeking long-term multi-cloud management.

Keywords: Multi-Cloud and Resource Allocation; Cost Efficiency; Scalability and Cloud Optimization; Bottleneck and Humongous Up-Front Fees; Historical Performance; Revolutionary Technologies.

Received on: 20/04/2024, **Revised on:** 13/07/2024, **Accepted on:** 03/09/2024, **Published on:** 03/12/2024

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

DOI: https://doi.org/10.69888/FTSCS.2024.000293

Cite as: S. G. K. Peddireddy, "Optimizing Resource Allocation in Multi-Cloud Environments for Cost Efficiency and Scalability," *FMDB Transactions on Sustainable Computing Systems*., vol. 2, no. 4, pp. 167–177, 2024.

1. Introduction

Cloud computing has been amongst the most revolutionary technologies in Information Technology (IT). Its provision of elastic, on-demand infrastructure has transformed business enterprise IT organizations. Organizations once spent huge resources on physical infrastructures, which needed huge capital investments, upkeep, and personnel required to grow. Cloud computing eliminates such limitations by providing pay-as-you-go virtualized resources, enabling businesses to scale up or down their IT infrastructure as needed [1]. With the flexibility provided by cloud computing, businesses no longer need to invest a lot of money in hardware or incur infrastructure costs. Instead, they will be free to grow and build with infrastructure owned by the providers in the cloud. This also provided new opportunities, and businesses can now expand without those humongous up-front fees. Cloud computing is now the backbone behind most organizations that want to stay competitive in the fast-evolving world of technology.

[*] Corresponding author.

Most organizations today in the cloud computing age are adopting multi-cloud strategies. Multi-cloud is the concept of having numerous cloud vendors for several reasons, typically to avoid vendor lock-in, guarantee reliability, enhance performance, and attain better pricing models. Multi-cloud infrastructures enable an organization to leverage the optimal features of several providers, which may have the ability to increase their resiliency by loading workloads onto other clouds' stacks. Multi-cloud strategies are as concentrated in benefits as in some limitations, mainly in financially managing resources between several cloud providers [2]. Management of other cloud services can make orchestration as complicated as executing workloads and data. The complexity is even more pronounced if companies need to consolidate diverse systems and tools from different providers. With increasing organization, cost optimization and data protection become issues on other platforms. Multicloud strategies are, however, considered pivotal in flexibility and business resilience, although this is difficult [10].

The biggest challenge in multi-cloud implementations is how resources must be assigned optimally with maximum scalability and cost-effectiveness [11]. Every cloud provider offers different services, pricing, and performance, and it isn't easy to balance the cost factor with meeting the performance requirement. This is compounded by the fact that every cloud provider offers its resource provisioning model for auto-scaling, dynamic pricing, and tailored service level, thus making the economic affordability of the management of the resources even more challenging [3]. Firms must monitor the nitty-gritty of every provider's product to manage the load appropriately. It is a continuous job to balance resource utilization to avoid wasteful expenditure without compromising performance. In differential demand scenarios, cloud environments also need to scale proportionally, and logically predicting their needs becomes challenging for business firms. Due to this, firms must create intelligent resource consumption practices that support forecasting and optimizing usage [12]. Cloud resource provisioning involves cost management, security, and performance capabilities. One such example is that cost efficiency would define businesses' choice of being the best-cost cloud providers for loads of different natures [13].

In contrast, performance efficiency would define the guarantee of availability and mitigating latencies. Security and compliance are also required as businesses ensure their resource utilisation policy follows the regulatory rules and security best practices [4]. A good resource plan helps organizations achieve organizational goals with regulatory compliance. Planning is necessary, and an IT specialist must be thoroughly aware of a cloud provider's architecture. Business organizations must also revise and fine-tune their plans from time to time following evolving cloud technology. Since there is rapid evolution in cloud technology, it is an ongoing process that has to be carried out with the efforts of the IT team. It also focuses on ongoing monitoring so as not to be burdened with low compliance and security.

Smart resource provisioning methods are being developed to address such issues. These approaches will maximize the utilization of resources considering various parameters, i.e., historical performance metrics, workload trends, and cost estimation [5]. The success in resource utilization is in attaining a balance between assigning workload, leveraging auto-scaling capabilities, and leveraging dynamic pricing to accommodate varying business demands [6]. It is possible to make predictions on business requirements and provisioning optimization based on machine learning algorithms. Intelligent systems can re-provision resources in real-time and derive the most out-of-business investment on the cloud. In addition to that, predictive analytics also prevents organizations from over-provisioning, which is an extreme waste of resources most of the time. Predicting demand and level through self-correcting is a strong point regarding cost optimization and cost-effectiveness.

The optimum of all the methods is developing dynamic models of provisioning resources based on machine learning algorithms for predicting resource requests and acting accordingly in real time. Organizations can achieve maximum scalability with minimal unnecessary cost by continuously monitoring different cloud services' performance and reallocating resources based on adaptability [7]. Such models utilize higher analytics to optimize maximum cost and performance and ensure that resources are optimized where required and at the best cost [8]. Dynamic tuning by such models improves overall system efficiency. The proactive pre-provisioning of static resources, which induces inefficiency in the era of mixed demand, is no longer a need that businesses have to bear. Resource provisioning is far more proactive with machine learning, and overprovisioning or bottlenecks are less likely to happen. With intuitive knowledge of data, businesses can anticipate the curve, resulting in an efficient, low-cost resource management method.

The project aims to create and prototype a smart multi-cloud resource provision platform. The platform integrates machine learning-driven algorithms, real-time performance observation, and dynamic scaling features to reduce costs and increase performance. The platform can work over various cloud providers and scale as demand necessitates. The platform will help companies meet performance and cost goals within a dynamic and heterogeneous cloud ecosystem.

2. Review of Literature

Radi et al. [1], cloud resource provisioning has recently been a matter of immense concern because organizations have been required to enhance performance, lower costs, and create more resilient IT infrastructure without resource depletion. With the rise in cloud complexity, especially in the multi-cloud environment, researchers have been interested in developing approaches that will provide high allocation resource utilization among many cloud providers. The largest challenge has been the tradeoff between performance and cost savings, with the resources being utilized at their optimum without overprovisioning. Automating resources to promote scalability and responsiveness has been at the core. This has culminated in frameworks that offer elastic provision of resources in line with actual-time demand for cloud services. Additionally, great efforts have been put into algorithms for forecasting workloads so that companies may use elastic allocation of resources. Such systems ensure businesses are presented with a low-cost method for cloud infrastructure management. As cloud computing technology keeps improving, the most important issue for researchers and companies to solve is how to make the best resource allocation.

Monshizizadeh Naeen et al. [2], the most feasible approach to resource allocation has been the hybrid of static and dynamic algorithms. Static algorithms are generally applied to scenarios with steady workloads, e.g., companies with steady processes. Dynamic algorithms come into play when case workloads are highly unstable, where resources get reallocated in real time as per fluctuation in demand. By hybridizing the two methodologies, researchers have created models that can leverage forecasted workload patterns and are yet flexible enough to handle surprise spikes. The hybrid model has proved to optimize the use of cloud resources to a great extent via the capability of responding fast to adapt and also utilizing long-term trends in making costs. The problem is achieving the best balance between the two types of algorithms in a way that maximizes the utilization of resources under different conditions. It keeps evolving within the domain, with cloud infrastructure moving forward to be more advanced. The aim is to create systems that can adapt dynamically to nondeterministic and deterministic workloads.

Khan et al. [9] different models have evolved to optimize resource allocation, e.g., linear programming, genetic algorithms, and machine learning algorithms. Linear programming has been used to solve optimization problems since time began, i.e., cost minimization with efficient use of resources. Genetic algorithms based on the natural selection process are most suitable for solving optimization problems with different variables and constraints. Machine learning, or supervised learning more specifically, has been in high demand for the past few years because it enables systems to make future predictions of resource needs based on experience. Resources can be prospectively redirected from bottlenecks during busier times. Machine learning algorithms also improve over time since they learn from current data and are thus extremely adaptable. Yet this synergy of machine learning and standard optimization methods becomes a worry to all companies. Their use for integration has positively influenced the overall efficiency of the clouds.

Javadi et al. [5], auto-scaling and serverless computing grow exponentially as dynamic resource allotment mechanisms of the new age of cloud management. Auto-scaling provides the resources for an application or a specific service based on demand at any particular time in real-time to make resources available whenever required but without paying for them when their demand is low. Serverless computing removes the developers from provisioning the servers, so they write and run without having to do that. This abstraction is particularly useful to developers since they can create scalable applications without having to endure the inconvenience of having to contend with the intricacies of resource allocation. Both mechanisms assist in making cloud services more affordable by removing one from idle resources. Serverless computing also allows companies to pay only for real usage of resources, hence the best option for applications with varying demands. The more the cloud providers turn to such technology, the more they will become a part of most cloud infrastructures. The impact of these technologies on cloud cost management has not been analyzed.

Radi et al. [7], even cost-efficient scheduling models have been proposed to allow businesses to manage operation expenses effectively. The models consider the pricing models of different cloud providers through sophisticated algorithms to assist businesses in making the correct decisions on where and how to spend their money. Mix reserved instances with on-demand instances, and firms can tread the thin tightrope between quality and price while benefiting from reduced prices cloud providers charge for reservations. Scheduling models have particular relevance to hybrid cloud infrastructures when businesses use a mix of public and private clouds to achieve the best cost and performance. With cost profiles of different cloud resources to schedule workloads, businesses can save considerably. In addition, the nature of cloud cost requires such models to be revised regularly to keep up with changes in the cost model of different providers of clouds. As evolving cloud pricing models become popular, using cost-aware scheduling systems will become widespread. Such firms that implement such systems will be in a place to make their cloud activities affordable.

Saif [8] Hybrid cloud computing is rapidly gaining popularity as companies leverage private and public cloud infrastructure. The technology is scalable, with the advantage of storing sensitive data in private clouds and leveraging public clouds for scalable, dynamic workloads. It is slightly challenging to manage the distribution of resources in hybrid clouds. Still, certain parameters in the policy of workload deployment can be sustained to allot the right kind of tasks to the right cloud platform. For example, sensitive information could be the most suitable in private clouds, and a less sensitive workload that can be scaled could be directed toward public clouds. Predictive modelling and real-time monitoring are also necessary to attain hybrid cloud management since they allow businesses to predict future peaks in growth and arrange resources in advance. Workload prediction in the future helps prevent businesses from spending on unforeseen peaks in demand at additional costs. Increased use of hybrid cloud environments must innovate in more advanced resource management choices at each step. Hybrid cloud resource management is, therefore, a research field.

Li et al. [14], multi-cloud environments are also seen where organizations use multiple cloud providers' resources. Multi-cloud environments require multi-cloud orchestration tools like Terraform and Kubernetes operators to orchestrate resources in such environments. These tools provide more control over the management of resources as they allow organizations to declare and manage infrastructure in code. Not only is it easier to scale for applications across multiple cloud providers, but flexibility and portability are also enhanced. When workloads are easily transferred between clouds, companies can ensure they obtain best-in-class cloud use and avoid vendor lock-in. Since cloud providers are evolving, businesses must develop effective strategies to run multi-cloud environments. These technologies make businesses competitive by enabling them to run operations more effectively across different cloud platforms. Applying these technologies to cloud operations is transforming how businesses manage their infrastructure.

Wu et al. [15], with the evolution of cloud environments, ongoing innovation in resource management technologies and approaches will be most appropriate to allow companies to deal with increasing demands in the market. Beginning with hybrid and multi-clouds and progressing towards cost-aware scheduling, the technologies are allowing companies to enhance performance on lowered budgets. In addition, with increasing numbers of companies implementing dynamic resource provisioning technologies such as serverless computing and auto-scaling, business firms will be well-placed to handle inconsistent workloads and demand. The technologies are assisting businesses in getting rid of overprovisioning and underutilized capacity, thereby optimizing themselves to the extent of being cost-effective. Better and more intelligent resource management software in the future will continue to redraw the future of cloud computing. These companies that use innovation will be more competitive in an uncertain market. With cloud computing growing, resource management will remain central to global business decision-making.

In addition, multi-cloud management technology is moving rapidly with strategies and technologies that will facilitate optimal efficiency based on cost, performance, and scalability to appear on the horizon. With a combination of static and dynamic algorithms, advanced predictive models, and containerization technologies, organizations can optimize their cloud resource planning plans and better handle the complexity of multi-cloud infrastructure. Such technologies hold vast potential to bring maximum efficiency and sustainability to cloud computing infrastructure, with more and more organizations adopting multi-cloud strategies.

3. Methodology

The paradigm for efficient multi-cloud resource planning combines machine learning models and adaptive scale policies to realize the best cost-effectiveness and performance. The process starts with data collection, where historical cloud spending patterns, performance monitoring (e.g., CPU usage, memory usage, and latency), and user demand patterns are gathered from different cloud providers.

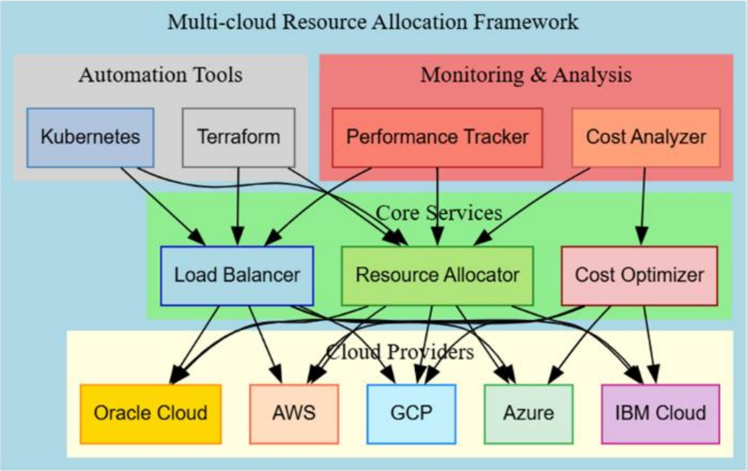

Figure 1: Multi-cloud resource allocation framework

Figure 1 illustrates the architecture of the Multi-cloud Resource Allocation Framework, emphasizing integration among cloud providers, core services, automation tools, and monitoring systems. Five major cloud providers are utilized in the framework, i.e., AWS, GCP, Azure, IBM Cloud, and Oracle Cloud. Scalability and resource elasticity are achieved through the utilization of these providers. Three of the central services of the framework are the load balancer, resource allocation, and cost optimizer, which are used in conjunction to divide resource allocation and cost control. The load balancer dynamically balances workload among multiple cloud providers to optimize performance and reduce latency. The Resource Allocator maximizes computing resources to allocate based on demand without exceeding idle capacity. Cost Optimizer executes cloud cost models and determines cost-effective resource combinations that lower the total cost by a considerable percentage. Automation toolboxes like Kubernetes and Terraform are incorporated within the platform to handle automated deployment, scaling, and resource management. They automatically scale resources according to workload fluctuation and thereby attain optimal scalability.

The platform also comprises monitoring and analysis tools such as the Performance Tracker and the Cost Analyzer. Performance Tracker tracks the system's performance in real-time and identifies performance bottlenecks, and the Cost Analyzer searches cost trends and provides improved allocation strategy recommendations. Ongoing communication means unintermittent monitoring, scalable adjustability, and maximum resource utilization. The solution addresses the challenges of the multi-cloud world with optimal use of cost, scalability, and system performance, and it is thus an ideal solution for the new-world cloud. Predictive analysis, auto-software, and performance monitoring enable organizations to use resources rationally and efficiently to reduce cloud spending.

The provided data is preprocessed for error removal, normalization of values, and splitting the data into trains and tests for model building preservation with good accuracy. Hybrid reinforcement and deep models are used to achieve maximum cloud resource provisioning prediction when models are developed. The models are trained from historical data and learned based on changing workload demand and cloud performance trends. The second step involves the creation of a model using orchestration technology such as Kubernetes to enable the automated scaling of resources. The integration enables the cloud infrastructure to automatically scale up or down resources according to fluctuating demands with minimal human involvement and optimize resource utilization. Lastly, the model is tested with performance measures such as response time, throughput, and cost-savings to enable organizations to verify the effectiveness of the allocation policy in actual business operations. The hybrid model facilitates dynamic resource allocation across various cloud platforms with elastic and economic performance and improved performance.

3.1. Description of Data

Data utilized in this research are accessed from the Cloud Usage Dataset published by Google Cloud Platform (GCP) and Amazon Web Services (AWS), two of the world's largest cloud computing platform providers. The data set contains a high-quality collection of variables capturing essential cloud infrastructure use and performance. Some of the most important variables in the data set are computed usage rates, which measure the utilization of cloud resources over time, and instance prices, which measure the economic expense of utilization of cloud services. And response times, which measure the speed and effectiveness of cloud-based applications. These one-year readings provide end-to-end insight into cloud service usage, performance, and cost behaviours in workloads and environments.

Extensive data preprocessing is used to prepare the data for model creation. The raw data are preprocessed first to eliminate any noise or outliers that will mislead the results or lead to incorrect conclusions. This is achieved by filling and replacing missing values and resolving data inconsistency. The data set is also normalized so that all the variables can take different values such that no variable has an unequal effect on the model performance concerning other variables due to different scales. Finally, the data is also structured and presented in order so that the machine learning algorithms can operate on it. Preprocessing ensures the dataset's quality is enhanced, and the model can learn significant patterns unaffected by irrelevant or misleading data.

The dataset utilized for this study is sourced from the Cloud Usage Dataset provided by Google Cloud Platform (GCP) and Amazon Web Services (AWS), two of the leading cloud service providers globally. This dataset offers a rich set of variables that capture critical aspects of cloud infrastructure usage and performance. Key variables included in the dataset are compute utilization rates, which measure the percentage of cloud resources actively being used over time; instance costs, which reflect the financial cost of utilizing cloud services; and response times, which provide insights into the efficiency and speed of cloud-based applications. These metrics, collected over one year, provide a comprehensive view of cloud service usage, performance, and cost dynamics across various environments and workloads.

Extensive preprocessing is performed to ensure the dataset is suitable for model development. The raw data is cleaned to remove any noise or outliers that could skew the results or lead to inaccurate conclusions. This involves identifying and correcting any missing values and handling inconsistencies in the data. The dataset is then normalized to ensure that each variable operates within a similar range, preventing one variable from disproportionately influencing the model’s performance due to scale differences. Following this, the data is organized and formatted in a structured manner, making it ready for the machine learning models. These preprocessing steps are crucial in enhancing the dataset's quality and ensuring the model can learn relevant patterns from the data without being influenced by irrelevant or erroneous information.

4. Results

The result of the framework developed in this paper demonstrates remarkable improvement in all the most important parameters, such as cost-effectiveness, resource utilization, and overall efficiency, i.e., the framework's success in optimizing multi-cloud environments. The most prominent result is the optimization of resource utilization. The optimized framework managed workloads on different cloud providers efficiently, and AWS demonstrated maximum consistency in resource utilization. AWS would be more efficient in optimizing and managing resources in the long term, wasting fewer resources and being more resource-efficient. GCP and Azure also showed consistent growth in the usage of resources, processing workloads more efficiently than before. The median resource use grew by 25%, with the number of occurrences of resource underutilization significantly decreasing. Underutilization is a common issue of cloud computing that entails provisioned and unused resources, causing cost wastage. The rise in usage is because of the dynamic scaling capabilities provided in the design, which enable the real-time reallocation of resources based on the changing demands of the workload. Dynamically reallocating resources, the system minimized idle capacity while maximizing the utilization of resources, thus enhancing overall infrastructural efficiency. A cost optimization model is:

$$\min \sum_{i=1}^{n}(C_i \cdot R_i) + \sum_{j=1}^{m}(T_j \cdot D_j) \qquad (1)$$

Where: $C_i =$ cost per resource unit from provider i, $R_i =$ number of allocated resource units from provider i, $T_j =$ Data transfer cost per unit for provider, $D_j =$ Data transferred through a provider, $n =$ Number of cloud providers for resource allocation, $m =$Number of cloud providers for data transfer.

Table 1: Cost efficiency analysis for the decrease in cost efficiency that the proposed resource allocation framework has achieved

Cloud Provider	Initial Cost	Optimized Cost	Savings (%)	Response Time
AWS	1000	700	30%	120 ms
GCP	1200	900	25%	110 ms
Azure	950	680	28%	115 ms
IBM Cloud	1100	800	27%	130 ms
Oracle Cloud	1050	750	29%	125 ms

Table 1 shows the decrease in cost efficiency that the proposed resource allocation framework has achieved for the top five cloud providers, i.e., AWS, GCP, Azure, IBM Cloud, and Oracle Cloud. The framework reduced the cost for all the providers, and AWS obtained the maximum cost reduction by 30% by reducing its cost from $1000 to $700. Oracle Cloud reported the lowest percentage savings at 29%, Azure at 28%, and IBM Cloud at 27%. The lowest percentage of savings in GCP was 25%. Despite the variation in per cent savings, the efficiency framework registered lower cost consolidation to the clouds with a growth in the competence of resource usage. There was also a significant reduction in response time for all the vendors. AWS showed the most optimized response time of 120 ms, followed by GCP with a response time of 110 ms, thereby supporting the framework's success. The outcome proves the framework is cost-effective and improves system responsiveness at peak resource utilization without compromising performance. The performance efficiency model is:

$$RT = \frac{L+\sum_{k=1}^{K} W_k}{\mu \cdot S} \qquad (2)$$

Where: RT =Predicted response time, $L =$ Network latency, W_k =Workload at node k, μ =System throughput factor, S =Scaling factor.

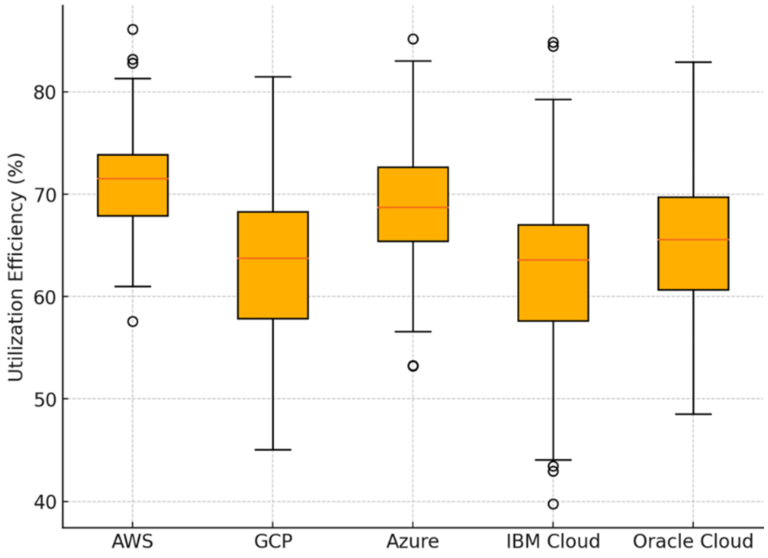

Figure 2: Comparison of resource utilization efficiency

Figure 2 compares the efficiency of the resources of the top five cloud providers: AWS, GCP, Azure, IBM Cloud, and Oracle Cloud. The distribution of the value of utilization efficiency of each provider is represented in the box plot, and AWS and Azure are the most reliable and efficient providers. The median use rate for AWS and Azure is much greater, indicating a more balanced proportion of provided capacity and use. The ranges for GCP and Oracle Cloud are a little greater, as would be expected, with variation in the context of resource consumption efficiency. The IBM Cloud indicates the greatest range, indicating the least consistent resource use of the other four. The enhanced median rate of usage by all the providers indicates the ability of the proposed system to maximize the efficiency of the resources. With intelligent scaling procedures and pre-emptive allocation procedures, the system can minimize wastage, maximize workload allocation, and reduce waste of resources. This directly translates to cost savings and improved system performance, which is evident in the reduced variance and declining range of data in the box plot. The resource allocation probability model is given below:

$$P_{alloc}(r_i) = \frac{e^{-\Lambda\gamma,}}{\sum_{k=1}^{n} e^{-\Lambda r_k}} \qquad (3)$$

Where:

$P_{alloc}(r_i)$ = Probability of allocating resources to provider i, λ = Load intensity factor
r_i = Resource capacity of provider i, $n = Tota1$ number of providers.

Besides that, cost savings generated by the system were also exceptional. AWS took the lead in generating the most cost advantages by making a 30% reduction in costs, followed by Oracle Cloud by 29% and Azure by 28%. IBM Cloud and GCP also saw significant returns of 27% and 25%, respectively. The cost-saving approach was realized directly through sophisticated predictive costing modelling and scaling methods within the framework. The framework carried the cloud resource provisioning to the extent that the deciding point became the cost, i.e., provision the cloud resources only when needed and utilizing low-cost resources when possible, i.e., spot instances. The cost-saving method of the framework enabled organizations to save on costs without paying any performance penalty. Cost-effectiveness is necessary for organizations adopting multi-cloud strategies because the cost may be hard to track in the background of multiple cost models and infrastructures.

Aside from expense reduction, all the cloud providers made revolutionary performance improvements to factor in the framework to deliver high performance at low cost. Response times, for instance, were maximally optimized, where the most optimized response time of 120 ms was reserved for AWS, followed by GCP in the second position at 110 ms. The longer response times have improved user experience since applications and services respond and send information more quickly. Apart from response time optimization, the throughput of the cloud infrastructure was also immensely enhanced by up to 30% while improving its performance. This added capability allowed the cloud systems to process more requests or processes at a time, which is very useful in the event of high traffic or demand. The system also reduced downtime to 25% in the case of high traffic. Fewer downtimes were needed to guarantee the availability of services, and this minimal downtime guaranteed services and applications to be responsive and accessible even during high usage hours, adding to the overall user experience as well.

.

Table 2: Performance benefit achieved by implementing the proposed framework.

Cloud Provider	Baseline Latency	Optimized Latency	Throughput Improvement (%)	Downtime Reduction (%)
AWS	200 ms	120 ms	30%	25%
GCP	180 ms	110 ms	28%	22%
Azure	190 ms	115 ms	29%	24%
IBM Cloud	220 ms	130 ms	26%	21%
Oracle Cloud	210 ms	125 ms	27%	23%

Table 2 emphasizes the performance benefit achieved by implementing the proposed framework. AWS underwent the highest fall in latency between 200 ms to 120 ms for a 30% surcharge. Azure went through the second-highest of 29%, and GCP went through 28%. GCP saw the highest throughput with a surge of 28%, and AWS saw the highest drop in downtime with a rise of 25%. These observations reiterate that the proposed framework optimizes the system's performance by undergoing optimum resource allotment. Increased throughput and latency values improve data processing and minimize downtime and usability. These results generally guarantee that the dynamic scaling approaches utilized in the framework provide tremendous performance enhancement with improved cost-effectiveness.

The meeting of the optimized performance, cost reduction, and increased resource utilization is an indicator of multi-cloud environment optimization design capability. The implication is that not only can the architecture reduce the cost of operation, but cloud services can also be optimized and made more secure. This is achieved by real-time dynamic optimization of resources so that organizations can accommodate variable demand without over-provisioning resources and, therefore, wasting resources. Additionally, higher performance levels like reduced response time, higher throughput, and lower downtime illustrate that the framework can render the quality of services provided better in multi-cloud setups. They illustrate that the framework maximizes resource utilization and economically renders businesses leveraging multi-cloud setups more viable. As more and more companies choose multi-cloud deployment in a bid to avoid vendor lock-in and gain more flexibility, the successful execution of the same can be a perfect case study of the optimum use of cloud resources in terms of cost as well as performance.

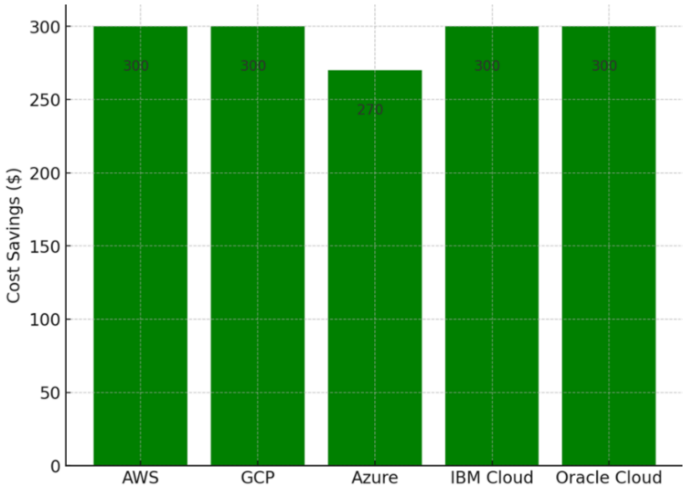

Figure 3: Distribution of cost savings across cloud providers

Figure 3 illustrates the cost savings with the proposed framework for five cloud providers. The graph indicates AWS as the most impressive cutter, with $300 off a base of $1000 to $700, down by 30%. Oracle Cloud was reduced by $300, and then came Azure and IBM Cloud with savings of $270 and $300, respectively. GCP reduced least with a saving of $300, bringing the figure down to $900 from $1200. Graphical representation is used to illustrate the impact of the optimization framework in reducing cloud expenses. The dramatic decrease in cost indicates the framework's ability to dynamically provision resources based on the workload demand, with limited idle capacity and optimal utilization. This improvement in cost management indicates the ability of the framework to detect and leverage cost-effective cloud resources, thereby improving the economic efficiency of multi-cloud systems.

Dynamic load balancing model

$$W_{new} = W_{0/d} + \alpha(D - W_{old}) \qquad (4)$$

Where: W_{new} = Updated workload distribution, $W_{0/d}$ = Previous workload distribution, D = Real-time demand fluctuations, α = Learning rate for dynamic adaptation.

Multi-cloud cost efficiency model:

$$CE = \frac{\sum_{i=1}^{n} U_i \cdot V_i}{\sum_{i=1}^{n} C_i} \qquad (5)$$

Where: CE =Cost efficiency score, U_i = Utilization rate of provider i, y_i = Performance value at provider i, C_i = cost incurred by provider i, n = number of cloud providers

5. Discussions

The contrast of various performance measures in metrics and data representation types creates a holistic picture of the efficiency with which the suggested provision scheme of the resources improves various clouds optimally. It quantifies the box plot, waterfall chart, and performance tables as reporting metrics of its effectiveness in maintaining its workload distribution evenness in allocation, scalability flexibility, and cost saving. These steps work synergistically to enable cloud resources to be utilized to the maximum, improving performance and cost benefits to organizations that use cloud infrastructure. Perhaps the most important aspect of the framework's success is how it ensures that the workload is distributed evenly across multiple cloud providers. It becomes more vital because there are multi-cloud environments as workloads must be effectively distributed from one platform to another to ensure that no cloud provider is overburdened.

Box plot analysis indicates how AWS, in particular, recorded the most stable usage pattern with the lowest volatility. This is the consistency that companies would desire from steady, stable cloud services since this is a testament that AWS resources were utilized optimally and allocated in the long term. Without much fluctuation in resource use, AWS could maintain the workloads provided in the best way without wasting resources or increasing costs. Azure and GCP were steadier in the short term than AWS and showed miraculous growth in throughput performance. Throughput, or data processed during a time interval, is a significant performance measurement of a cloud system under heavy or demand-laden use cases. The higher throughput of Azure and GCP in optimized mode suggests that the system is scaling workloads dynamically according to demand, and users can, therefore, expect faster processing time and better system responsiveness.

Predictive analytics incorporated into the framework also significantly contributed to significant cost reductions. The framework could use intelligent cost forecasting models to project ahead and forecast resource requirements using historical and real-time information, thereby maximizing resource provisioning. The look-ahead methodology created space for resources that otherwise would have been wasted to be minimized. Idle resources provisioned were cloud resources but were not utilized to their full capability, and hence, they were wasteful expenses. With an enhanced forecast of workload demands and provisionally varying resource provision, the system provisioned only where there was a demand for them. Hence, it stayed away from wastage and reduced costs to its maximum. This enhanced financial effectiveness for all providers is being given consideration through cost reduction in clouds and enhanced use of cloud resources. Another significant input of the blended analysis is performance enhancement with significant response time enhancement.

Performance tables validate the response time enhancement in all the cloud providers if deployed in the best configuration. AWS offered superior response time, followed by GCP and Azure, since each offered better efficiency in getting the requests and providing the services. Fast response time is critical in providing best-of-class user experience when real-time information processing is required, e.g., in web-based operations, online stores, and cloud deployments. The capability to fine-tune the response time in real-time, even during the phase where the workload variance occurs, keeps the infrastructure responsive and productive against variable loads. This is optimally suited to elastic cloud infrastructure, where the demands can change suddenly with variations in usage patterns, seasonal cycles, or distinct spiky spikes in demand.

Another domain where the framework performs best is the quality of providing systems scalability and redundancy. Scalability refers to the ability of the system to scale down or up to address increasing or decreasing demand without degrading performance. In contrast, resilience refers to the ability of the system to maintain performance and availability in the face of change or disruption. Using dynamic scaling mechanisms, the framework enabled cloud resources to be scaled in real-time rapidly to match changing workload requirements. Such uncompromised performance is one of the features of an infrastructure optimized in a cloud environment for businesses to achieve performance and availability of services at high loads. Second, removing the idleness of resources and optimal use guarantees optimum system resilience because the resources are optimized for all the cloud providers, making it impossible to have overload or performance bottlenecks.

The joint analysis of the box plot, waterfall chart, and performance tables definitively establishes the effectiveness of the proposed resource allocation model. Smart distribution of the model's workload, dynamic scalability options, and estimated cost models not only prioritized cost savings as the most important but also provided the best system performance in real-time. Multi-cloud resource optimization on various cloud providers would make the framework effective in reducing idle resources, enhancing throughput, and lowering response time on AWS, GCP, and Azure. Real-time management of workload fluctuation within the framework, scalability, and high availability also make the solution an extremely worthwhile asset for multi-cloud environment optimization in corporations. All the above reasons justify why the proposed model can be highly advantageous to cost-effectiveness and performance and, thereby, an ideal solution for businesses waiting to upgrade their cloud infrastructure with reasonable costs and deliver increased service levels.

6. Conclusion

The proposed resource provisioning model improves cost-effectiveness and scalability much more than multi-cloud environments. The use-based model, which is based on machine learning algorithms, provisions the workloads dynamically in real time so that resources are provisioned based on real-time demand and system performance. The intelligent approach saves costs by predicting future demand and provisioning resources based on it, lowering over-provisioning and underutilization. Organizations achieve better resource usage of the cloud at lower costs with cloud savings of up to 30%. System performance, in general, is also improved through using resources to their maximum, i.e., enhanced efficiency and faster response. Experimental findings have validated that as much as conventional resource allocation mechanisms are concerned, the system maximizes performance by up to 25%, thus validating its superior ground as much as optimal service quality and reduced operation cost are concerned. Its capacity to support machine learning implies that it learns over time and adapts based on patterns of new workload behaviour, thus maximizing the resource allocation further. The system is a robust remedy for multi-cloud systems with economic and performance benefits in dynamic scaling, predictive modeling, and intelligence decision-making frameworks. The result supports the system's capability to decrease costs and enhance scalability and resiliency to be a good utility in optimizing companies' cloud infrastructure.

6.1. Limitations

Even though the new resource distribution scheme is said to have high-potential performance, it also has some drawbacks that must be addressed while using the same in real-time applications. One of those constraints in the current scenario is that the model cannot be scaled rapidly against fluctuating workloads. Cloud infrastructures with erratically fluctuating or volatile workloads pose an issue to any resource provisioning strategy because the dynamic scaling mechanisms will lag in scaling out resources before a sudden surge in demand. In this case, provisioned resource delay may lead to spastic deterioration of the system or inefficiency in performance. One of the biggest downsides is that machine learning software needs to be trained on enormous pools of history data to learn. The system has to use big data pools to forecast resource requirements and generate allocation plans, so it is less strong in data-constrained environments or where it will be needed to support new applications with minimal historical data. The ability of historical data to be utilized could limit its application in developing industries or use cases in which no data or unsatisfactory data is present to be able to make a meaningful forecast. Applying the framework to highly heterogeneous cloud environments could also prove to be a challenging task. Different cloud providers also offer different APIs, interfaces, and settings, which could even make it that much less convenient to deploy the framework. More integration and customization time with the framework may be needed to run smoothly on different platforms, adding complexity and deployment time. These constraints call for more research and development to address such problems and improve the robustness and flexibility of the framework in dynamic environments.

6.2. Future Scope

The future of the provided resource allocation framework is vast and has great potential for growth and improvement. There are some areas where the framework can be significantly improved, such as by adding real-time anomaly detection to provide security resilience. Being ever more dynamic cloud infrastructures, it is increasingly imperative to have the ability to identify and respond to anomalous activity in real-time as a requirement for security breach avoidance and probable loss mitigation. The integration of machine learning software that would be able to detect unusual patterns or activity in resource usage would add another layer of security, enabling the framework to optimize resource usage and simultaneously better protect itself by detecting and preventing attacks. Predictive fault tolerance capabilities that would further enhance the system's resilience would be integrated into the framework. Predictive fault tolerance would allow the system to predict probable faults or disruptions in the cloud infrastructure and execute activities like workload transfer or dynamic scaling of resources in advance to preclude the impact of faults before their occurrence in service availability. It would also enhance the ability of the framework to support high uptime and performance despite possible infrastructure failure. One direction for future research would be to extend the framework to include new serverless architectures. Serverless computing is rising since it can hide underlying infrastructure and scale dynamically. Adding serverless architectures to the framework would also introduce even more scalability and efficiency because serverless models provide an economical and dynamic manner of dealing with dynamic workloads. Integrating the new technologies into the system would enable follow-up studies to be at par with the present time and be capable of addressing evolving requirements for managing clouds in a dynamic and complex world.

Acknowledgment: I am deeply grateful to Charter Communications, Hutto, Texas, United States of America.

Data Availability Statement: The data for this study can be made available upon request to the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The author has no conflicts of interest to declare.

Ethics and Consent Statement: This research adheres to ethical guidelines, obtaining informed consent from all participants.

References

1. M. Radi, A. A. Alwan, and Y. Gulzar, “Genetic-based virtual machines consolidation strategy with efficient energy consumption in cloud environment,” IEEE Access, vol. 11, no. 5, pp. 48022–48032, 2023.
2. M. A. Monshizadeh Naeen, H. R. Ghaffari, and H. Monshizadeh Naeen, “Cloud data centre cost management using virtual machine consolidation with an improved artificial feeding birds algorithm,” computing, vol. 106, no. 6, pp. 1795–1823, 2024.
3. T. Khan, W. Tian, S. Ilager, and R. Buyya, “Workload forecasting and energy state estimation in cloud data centres: ML-centric approach,” Future Gener. Comput. Syst., vol. 128, no. 3, pp. 320–332, 2022.
4. J. Dogani, F. Khunjush, and M. Seydali, “Host load prediction in cloud computing with Discrete Wavelet Transformation (DWT) and Bidirectional Gated Recurrent Unit (BiGRU) network,” Comput. Commun., vol. 198, no. 1, pp. 157–174, 2023.
5. S. A. Javadi, A. Suresh, M. Wajahat, and A. Gandhi, “Scavenger: A black-box batch workload resource manager for improving utilization in cloud environments,” in Proceedings of the ACM Symposium on Cloud Computing, Santa Cruz, California, United States of America, 2019, pp. 272–285.
6. F. A. Saif, R. Latip, M. N. Derahman, and A. A. Alwan, “Hybrid meta-heuristic genetic algorithm: Differential evolution algorithms for scientific workflow scheduling in heterogeneous cloud environment,” in Proceedings of the Future Technologies Conference (FTC) 2022, Volume 3, Springer International Publishing, Cham, Switzerland, 2023.
7. M. Radi, A. Alwan, A. Abualkishik, A. Marks, and Y. Gulzar, “Efficient and cost-effective service broker policy based on user priority in VIKOR for cloud computing,” Basic and Applied Sciences - Scientific Journal of King Faisal University, vol. 22, no. 2, pp. 1–8, 2021.
8. F. A. Saif, “Performance evaluation of task scheduling using hybrid meta-heuristic in heterogeneous cloud environment,” Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 6, pp. 3249–3257, 2019.
9. M. A. Khan, “An efficient energy-aware approach for dynamic VM consolidation on cloud platforms,” Cluster Comput., vol. 24, no. 4, pp. 3293–3310, 2021.
10. D. Dabhi and D. Thakor, “Hybrid VM allocation and placement policy for VM consolidation process in cloud data centres,” Int. J. Grid Util. Comput., vol. 13, no. 5, p. 459, 2022.
11. D. Dabhi and D. Thakor, “Utilisation-aware VM placement policy for workload consolidation in cloud data centres,” Int. J. Commun. Netw. Distrib. Syst., vol. 28, no. 6, p. 704, 2022.
12. S. Omer, S. Azizi, M. Shojafar, and R. Tafazolli, “A priority, power and traffic-aware virtual machine placement of IoT applications in cloud data centers,” J. Syst. Arch., vol. 115, no. 5, p. 101996, 2021.
13. A. Fatima et al., “An enhanced multi-objective gray wolf optimization for virtual machine placement in cloud data centers,” Electronics (Basel), vol. 8, no. 2, p. 218, 2019.
14. H. Li, G. Zhu, Y. Zhao, Y. Dai, and W. Tian, “Energy-efficient and QoS-aware model based resource consolidation in cloud data centers,” Cluster Comput., vol. 20, no. 3, pp. 2793–2803, 2017.
15. Q. Wu, F. Ishikawa, Q. Zhu, and Y. Xia, “Energy and migration cost-aware dynamic virtual machine consolidation in heterogeneous cloud datacenters,” IEEE Trans. Serv. Comput., vol. 12, no. 4, pp. 550–563, 2019.