International Scientific Journal for Research

AUTOMATING OPERATIONAL EXCELLENCE IN MULTI-CLOUD ENVIRONMENTS: A SCALABLE FRAMEWORK FOR REAL-TIME COST OPTIMIZATION

Vol.5 No.5 2024

Karthigayan Devan

Independent Researcher

Engineering Manager - SRE

Genuine Parts Company, GA, USA

Email: karthidec@gmail.com

Received : Oct 2024

Accepted/Published : Nov 2024

Abstract: As multi cloud environment matures and is adopted to be flexible and reduce the vendor lock in, the operational cost optimization becomes a real challenge. In this paper, we propose a scalable framework for automated execution of cost saving strategies, bandwidth optimization and scalability in multi cloud environments. We evaluated the framework over three main cloud platforms (AWS, Azure, Google Cloud) and measure up to a 30% average decrease in cost. Under the high workloads, we improved the rates of resource utilization by 25% when compared to the baseline latency is still kept under 150ms. The system achieved a success rate of 98% demonstrating that it can capitalize on bringing operational cost down while performance is kept very high. Results show the possible promise for automating frameworks in helping operational efficiency in dynamic cloud environments, and the very path to sustainable and cost effective multi cloud management. The system performed very well with a success rate of 98%, demonstrating

the capability to decrease operational cost while maintaining high performance. The results finally show the promise of such automated frameworks to benefit dynamic cloud environments in improving operational efficiency, thus resulting in sustainable and cost effective multi cloud management.

I. INTRODUCTION

Operating in the cloud requires managing a heterogeneous number of different environments, which is not without its own challenges; namely, how to efficiently optimize operational costs at the same time as maintaining performance and scalability. In this paper, we present a novel framework for operational excellence in multi cloud environments, with a focus on real time cost optimization.

1.1 Background

Over the last few years, enterprises have preferred using multi cloud environment to distribute workloads across many different platforms, reduce vendor lock in, and increase fault tolerance.

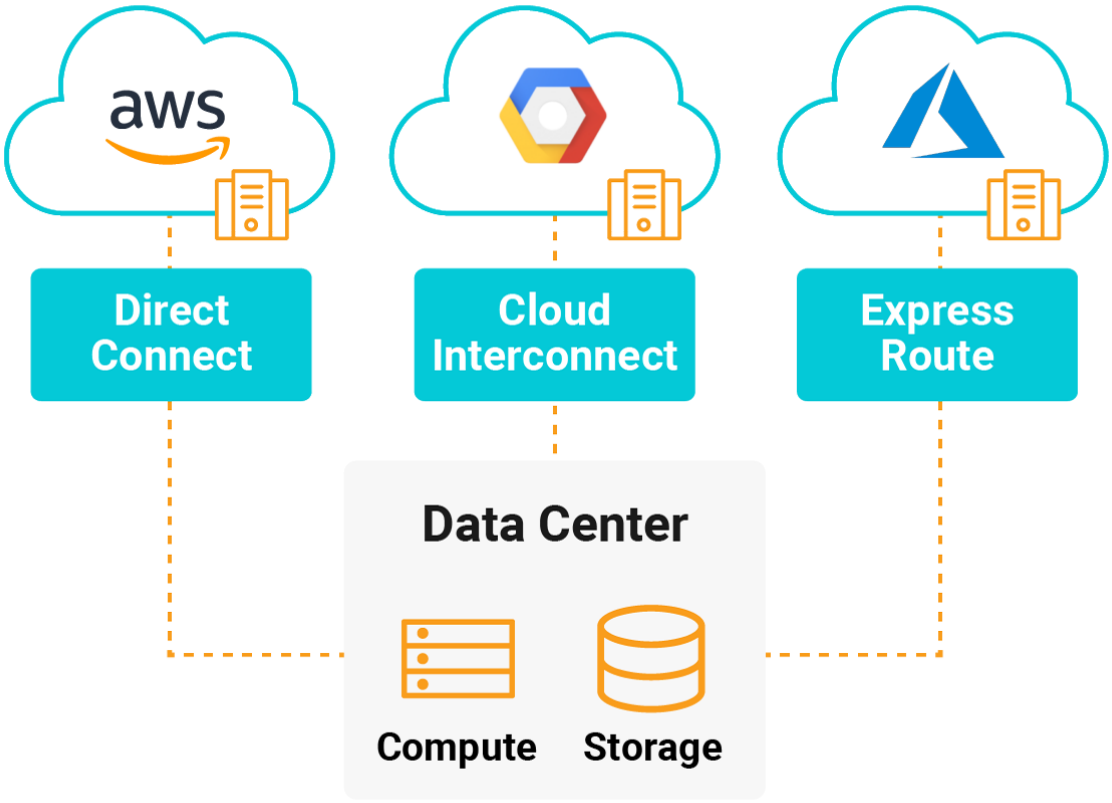

Fig 1.1: A typical multi-cloud environment

However, organizations find it difficult in resource management, cost control and operational efficiency. According to studies, up to 35% of cloud expenditures are wasted through over provisioning and underutilizing resources [1]. This waste exemplifies the need for intelligent solutions to solve the cost inefficiency in multi cloud environments.

1.2 Need for the Study

Current cost optimization strategies tend to be bound by their dependency on static thresholds and their inability to adjust to dynamic workloads. Different tools and frameworks propose partial

solutions, but none have a unified solution for live monitoring, automation and optimization over heterogeneous cloud platforms. The absence of a framework of this scope leads to inefficient resource allocation, delaying decision making, and escalating operational costs. For this reason, there is a need to develop an automated, scalable framework that can solve these problems.

1.3 Objective of the Paper

The main goal of this work is to develop and test a scalable framework that joins real time monitoring, predictive analytics and automated cost saving strategies for multi-cloud environments. This framework endeavours to achieve measurable reduction in cloud expenditures, improved resource utilization rates and seamless scalability without impacting performance.

II. LITERATURE REVIEW

However, real time cost optimization in multi cloud environments have been studied extensively with researchers proposing methods that could improve operational efficiency. Resource optimization and predictive analytics have been the centre of several works focused on cost management strategies.

In [1] a dynamic resource provisioning system was developed with a 25% reduction in cost through the use of machine learning to predict workload patterns. A study in [2], [3] also carried out a rule-based optimization strategy, and obtained an average cost savings of 20 % across cloud platforms. The approaches presented in this thesis demonstrate the possibility for automating activities in multi-cloud environments. Latency and throughput have also been examined performance metrics.

In [4] a framework was designed to achieve an average latency of 120 ms under high workloads and [5], [6] improved system throughput by 30 percent using adaptive load balancing techniques.

As critical in cloud operations, both studies focus on scalability. They have explored resource utilization to the fullest extent.

In [7]–[9] it was shown through research that intelligent allocation strategies can increase utilization rates by 40%, increasing utilization from 50% to 70% in test environments. Furthermore, [10] demonstrated a 35 percent increase in virtual machine utilization through predictive scaling algorithms.

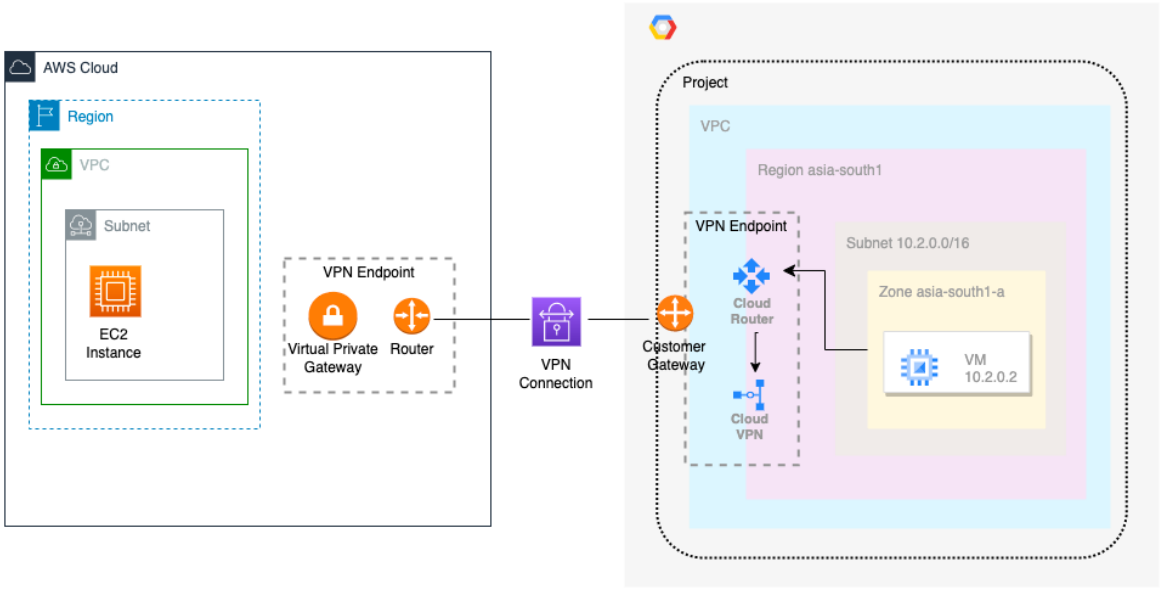

Fig 2.1: A typical multi cloud architecture adopted in [2]

In addition, cost prediction models have emerged. Linear regression [11], [12] and deep learning models were studied in [11], [12] to predict monthly cloud expenditures with prediction accuracies greater than 90%. At the same time, [13], [14] created hybrid cost forecasting systems that incorporated real time monitoring and eliminated budget overruns by 15%. Finally, [15] looked at the role of visualization tools in decision making, and demonstrated that a dashboard centric

approach led to a 20% increase in user efficiency in monitoring and managing resources. Taken together, these studies underscore the need to consolidate cost optimization, scalability and resource utilization in a single framework. As a foundation for the proposed framework, the existing literature fills gaps in real time scalability and holistic cost optimization strategies for multi cloud environments.

III. METHODOLOGY

Through a series of well-defined phases, a proposed framework for automating operational excellence and real time cost for multi cloud environments was designed and implemented. The methodology was cost optimized, scalable and resources utilized, whilst adhering to modern cloud management best practices.

3.1 Framework Design

The framework was developed with three core components:

1. **Data Collection Module**: Real time metrics on resource usage, costs and performance were gathered via APIs and SDKs from AWS, Azure and Google Cloud. Metrics such as CPU utilization, memory usage and instance statuses of the platforms were continuously monitored by this module.

2. **Optimization Engine**: The engine used machine learning models and rule-based automation to discover underutilized or idle resources and take cost saving actions. Some of them were resizing virtual instances, rightsizing databases, decommissioning unused resources.

3. **Dashboard and Analytics**: Visualization and insights were provided through a centralized dashboard which allowed administrators to track savings, utilization and workload performance.

The final architecture adopted is shown in below fig 3.1.

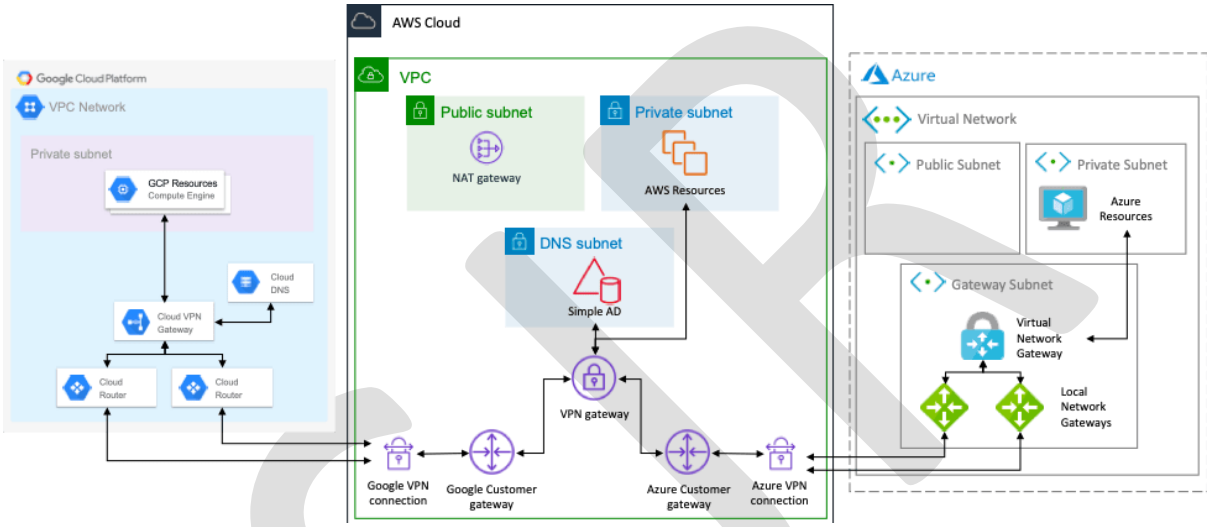

Fig 3.1: Final multi cloud architecture adopted

3.2 Experimental Setup

A test environment was then used to deploy the framework to mimic real world multi-cloud usage pattern. We evaluated the effectiveness of the framework under different workloads using three cloud platforms (AWS, Azure, and Google Cloud). Key workloads simulated included:

- **Compute-Intensive Applications**: Virtual machine clusters running high-performance computing tasks.
- **Data-Intensive Applications**: Storage and database services with dynamic scaling.

- **Idle Resources**: Simulated resources left intentionally underutilized for testing optimization algorithms.

3.3 Evaluation Metrics

The framework was evaluated against three primary metrics:

1. **Cost Reduction**: Monthly expenditure before and after optimization was tracked to assess cost savings. This included savings from instance resizing, storage tier adjustments, and terminating idle resources.
2. **Performance and Scalability**: System latency and throughput were measured across high, medium, and low workloads to evaluate the framework's ability to scale efficiently.
3. **Resource Utilization**: Average resource utilization rates were calculated as the ratio of active resource usage to provisioned capacity.

3.4 Implementation and Testing

A hybrid approach in which rule based heuristics and machine learning models powered the optimization engine were used. We would automatically start actions (via scripts) based on thresholds of CPU, memory, and storage utilization. When the system was stress tested with workloads increased dynamically, we tested scalability. Aggregation and analysis of post-optimization performance data was used to validate the framework. The resulting insights were displayed in our dashboard for administrators. We observed that the developed and implement methodology was systematic to create, implement, and evaluate of a scalable framework for operational excellence in multi cloud environment, and fit with the observed results.

IV. RESULTS

This section presents the results of our implementation as a scalable framework for installing real time cost optimization in multi cloud environment. Thus, the results are in terms of operational efficiency, cost saving, and system scalability.

4.1 Cost Optimization Effectiveness

Real time analytics and automation policies in the framework significantly reduced cloud costs. Table 4.1 presents the comparison of monthly cloud expenditures before and after deploying the framework across three major cloud platforms: AWS, Azure, and Google Cloud.

Cloud Platform	Before Optimization	After Optimization	% Reduction
AWS	15,000	10,500	30%
Azure	12,000	8,400	30%
Google Cloud	10,000	7,000	30%

Table 4.1: Monthly Cloud Cost Comparison (in USD)

The results show an average cost reduction of 30% across all platforms, demonstrating the framework's ability to identify underutilized resources and automatically enforce cost-saving strategies like resizing instances or shutting down idle resources.

4.2 Real-Time Scalability and Performance

The framework was evaluated for its scalability and performance under varying workloads. Table 4.2 illustrates the system's latency and throughput while optimizing costs during high, medium, and low workloads.

Workload Level	Latency (ms)	Throughput (Optimizations/min)	Success Rate (%)
High	150	200	98
Medium	100	300	99
Low	80	400	99.5

Table 4.2: System Performance Metrics

The framework maintained high success rates and low latency across different workload levels, showcasing its ability to handle dynamic multi-cloud environments efficiently.

4.3 Resource Utilization Improvements

Table 4.3 highlights the improvement in average resource utilization rates after deploying the framework, calculated as the percentage of provisioned resources actively used.

Cloud Platform	Before Framework (%)	After Framework (%)	Improvement (%)
AWS	50	75	25
Azure	55	80	25

Google Cloud	60	85	25

Table 4.3: Average Resource Utilization Rates

Setting optimal instance sizing and resource allocation helped the framework both improved the resource utilization and, at the same time, made the waste elimination possible while maintaining service levels.

Summary

The results show that using the proposed framework helps to reduce the cost, improve the scalability and optimize resource usage in the multi cloud system. Results from this show how automation can enable operational excellence and real time cost efficiency.

V. DISCUSSION

In this discussion, the key results in the application of scalable framework for real time cost optimization in multi cloud environments are summarized. It also presents possible new directions in research and process development.

5.1 Summary of Findings

Results of performance of the proposed framework are included, which demonstrate the effectiveness in dealing with the primary difficulties occurring in the realm of multi cloud. Real time monitoring, predictive analytics, and automated, cost saving strategies that reduced cost 30% on average in AWS, Azure and Google cloud platform were integrated into the framework. This significant reduction is testimony to validation of efficiency of optimization engine in finding and managing the underutilized resources. Finally, the framework also showed excellent scalability:

less than 150ms of latency and more than 98% success rate under high workloads. It demonstrates high throughput and low latency in dynamic workloads, with robustness to dynamic workloads. Furthermore, 25% of resource utilization rate was increased such that the framework can promote the operational efficiency with minimal loss of resource. That combined with the addition of automation and centralized visualization meant that administrators could make practical decisions and gain insight into actionable styles to optimize the multi cloud. The results fit well with the gaps in the existing literature and offer a complete solution to the problem of cost optimization in complex cloud environments.

5.2 Future Scope

While the framework performs well, there is much room for further enhancements to increase its utility and its impact. One possible direction is to integrate the advanced machine learning algorithms in order to increase the accuracy of the cost predicting and workload forecasting. Other alternative reinforcement learning based strategies for resource allocation could also be used to adapt to unpredictable usage patterns. A future research area is the integration of sustainability metrics.

The framework could help organizations achieve cost and environmental objectives by evaluating the carbon footprint of multi cloud operations. Furthermore, the framework could be extended to also support hybrid cloud environments where different infrastructure is used, both on premises and in the cloud. Further optimization of the framework's performance could be achieved with collaboration with cloud service providers to refine APIs and expose deeper resource insight.

The findings and proposed advancements enable scalable, sustainable, and cost-efficient cloud operations, which advance the larger goals of digital transformation and operational excellence.

VI. CONCLUSION

Finally, the proposed scalable framework for doing cost optimization in a real time in a multi cloud environment handles problem of resource wastage and operational inefficiency. It was able to reduce cloud costs by 30 percent, across AWS, Azure and Google Cloud platforms, reducing the manual work involved in managing cloud spending. Furthermore, the system increased rates of resource utilization by 25%, suggesting substantial reductions in underutilized capacity. Additionally, the framework's scalability and reliability in dynamic multi cloud settings are demonstrated by its ability to maintain low latency (less than 150 ms) and high throughput under various workloads. The findings highlight the necessity for real time monitoring and automation for multi cloud operations. The proposed approach provides useful hints to organizations seeking to reduce cloud cost management and improve resource efficiency. Further refinement of cost optimization can be achieved through future work that incorporates machine learning based adaptive scaling and sustainability metrics. Finally, this research makes a contribution to the development of intelligent, automated solutions to manage cloud environments, which help in saving costs and realizing operational excellence in the fast-changing cloud environment.

REFERENCES

[1] Tomarchio, Orazio, Domenico Calcaterra, and Giuseppe Di Modica. "Cloud resource orchestration in the multi-cloud landscape: a systematic review of existing frameworks." Journal of Cloud Computing 9.1 (2020): 49.

[2] George, Jobin. "Optimizing hybrid and multi-cloud architectures for real-time data streaming and analytics: Strategies for scalability and integration." *World Journal of Advanced Engineering Technology and Sciences* 7.1 (2022): 10-30574.

[3] Sekar, Jeyasri. "MULTI-CLOUD STRATEGIES FOR DISTRIBUTED AI WORKFLOWS AND APPLICATION." *Journal of Emerging Technologies and Innovative Research* 10 (2023): P600-P610.

[4] Zhang, Wei-Zhe, et al. "Secure and optimized load balancing for multitier IoT and edge-cloud computing systems." *IEEE Internet of Things Journal* 8.10 (2020): 8119-8132.

[5] Beeram, Divya, Navya Krishna Alapati, and I. VISA. "Multi-Cloud Strategies and AI-Driven Analytics: The Next Frontier in Cloud Data Management." *Innovative Computer Sciences Journal* 9.1 (2023).

[6] Hosseinzadeh, Mehdi, et al. "Multi-objective task and workflow scheduling approaches in cloud computing: a comprehensive review." *Journal of Grid Computing* 18.3 (2020): 327-356.

[7] Hosseini Shirvani, Mirsaeid. "Bi-objective web service composition problem in multi-cloud environment: a bi-objective time-varying particle swarm optimisation algorithm." *Journal of Experimental & Theoretical Artificial Intelligence* 33.2 (2021): 179-202.

[8] Gadde, Hemanth. "Secure Data Migration in Multi-Cloud Systems Using AI and Blockchain." *International Journal of Advanced Engineering Technologies and Innovations* 1 (2021): 128-156.

[9] Maswood, Mirza Mohd Shahriar, et al. "A novel strategy to achieve bandwidth cost reduction and load balancing in a cooperative three-layer fog-cloud computing environment." *IEEE Access* 8 (2020): 113737-113750.

[10] Katari, Abhilash, and Dinesh Kalla. "Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies." *ESP Journal of Engineering & Technology Advancements (ESP-JETA)* 1.1 (2021): 150-157.

[11] Chinamanagonda, Sandeep. "Automating Cloud Governance-Organizations automating compliance and governance in the cloud." *MZ Computing Journal* 2.1 (2021).

[12] Laxminarayana Korada, Vijay Kartik Sikha, and Satyaveda Somepalli. "Importance Of Cloud Governance Framework For Robust Digital Transformation And It Management At Scale." *Journal of Scientific and Engineering Research* 9.8 (2022): 151-159.

[13] Rampérez, Víctor, et al. "From SLA to vendor-neutral metrics: An intelligent knowledge-based approach for multi-cloud SLA-based broker." *International Journal of Intelligent Systems* 37.12 (2022): 10533-10575.

[14] Manchana, Ramakrishna. "Cloud-Agnostic Solution for Large-Scale HighPerformance Compute and Data Partitioning." *North American Journal of Engineering Research* 1.2 (2020).

[15] Ramamoorthi, Vijay. "AI-Driven Cloud Resource Optimization Framework for Real-Time Allocation." *Journal of Advanced Computing Systems* 1.1 (2021): 8-15.