International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 10 | Oct – 2025 DOI: 10.55041/ISJEM05070
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Cloud Cost Efficiency: Strategies and Implementation

Akshay Jha

Indore, India

I. ABSTRACT

Clouds of resources manifest on demand, with every commodity charged per usage, but operational costs continue to skyrocket for organizations due to idle resources, over-provisioning, and improper workload management.

This thesis undertakes cloud-costing optimization and embraces strategies for efficient resource utilizations, automation, and predictive monitoring.

These are instance right-sizing, reserved, and spot instances, automated scheduling, serverless computing, and storage lifecycle management with networking optimization.

This research puts even more emphasis on FinOps practices as financial accountability and machine learning methodologies that aim to forecast resource usage for a more proactive cost control procedure.

A prototype framework facilitating reduction across major cloud vendors (AWS, Azure, GCP) via Infrastructure as Code and monitoring tools is built and evaluated.

From the experiments, results show that such an optimization can reduce costs by 30-40% while preserving performance and service-level agreements.

This research offered a comprehensive model towards sustainable and efficient cloud cost management.

II. INTRODUCTION

In the current digital paradigm, cloud computing is a force shaping transformations and innovation as it enabled businesses from the rigidity of inflexible, capital-intensive infrastructural models onto the scalable, on-demand meaning of services.

It is offered by the cloud so that companies can access and use compute resource things like virtual machines, storage and networking capabilities without worrying about owning any physical infrastructure and maintaining it.

The cloud platforms, including AWS, Azure, and GCP, have been drivers of great innovation during their lifespan due to their flexibility, global reach, and more extensive service portfolios.

Scholars sometimes indicate that the magical nimble and cheap nature of clouds ends with the tag line or the next website advertisement; generally, this is not perceived first-hand. Instead, an enterprise will notice hidden costs, costs running away, and challenges in managing usage effectively.

It is reported that, on several occasions, over 30% of expenditure is wasted on cloud resources because they are purchased but are not used, over-provisioned, and absolutely without governance.

As the move into the cloud increases, therefore, the cloud expenditure needs to be optimized so that businesses continue having a margin to work on and thereby stand the test of time.

1. Key Features of Cloud Computing

a) **"On Demand Self Service"-** The user can automatically provision undesired computing resources like processing power, storage or applications without manual intervention necessitated from the respective service providers.

b) **"Broad Network Access-** The cloud service are connected upon the network and accessed through the standard mechanisms, into a framework enabling its use on laptops, mobile phones, tablets, and the likes".

c) **Resource Pooling-** The liquidity or computing resources are implemented in a pool setup using shared infrastructure model where both physical and

virtual resources are dynamically assigned based on customer demand, aligning limitation, and flexibility.

d) **Rapid Elasticity-** The resources evolve in time following the increase and reduction of workload, most often automatically, with the end-user perceiving such an available capacity as limitless.

e) **Measured Service-** Using a capacity of metering, the cloud system administers and optimizes resource usage in an automated manner.

Such provides transparency of billing and assists in matching costs with consumption.

These benefits create an appealing description of the cloud for corporations: agility and scalable lack of alignment between business needs and IT costs.

Yet, at times, though the cloud is meant to be efficient and save money, things become complicated.

Hidden fees, cost overruns, and inefficient management of usage are some of the challenges enterprises faces.

Studies have found that almost one-third of cloud budgets may go to waste due to unused resources, overprovisioning, or lack of governance.

So as cloud use increases, every company needs to maintain cloud cost optimization to remain profitable and thrive in the long term.

III. PROBLEM STATEMENT

The general financial uncertainty has been introduced through the shift to cloud environments.

While the pay-as-you-go skills keep all capital expenses out, uncontrolled bills appear when resources go mismanaged or underutilized.

Many organizations then struggle to watch their consumption, parse through complex billing forms, and try to enforce cost-control mechanisms.

With all this lacking in transparency and efficiency, operations become more expensive, and resources are taken from innovation.

This brings us to the problem resolved in this thesis: inefficient utilization of cloud resources and resultant financial wastage.

Since organizations find it hard to bring cloud expenses into alignment with business objectives without implementing systematic strategies and frameworks, the stage is set for approaches that seek to optimize cloud costs while not degrading system performance, scalability, and reliability.

IV. OBJECTIVES OF THE STUDY

1. To analyze key factors contributing to a high cloud expenditure.
2. To study and evaluate existing cloud cost optimization strategies.
3. To design a framework that integrates cost-efficiency practices into cloud operations.
4. To implement and test selected optimization techniques in a real or simulated environment.
5. To measure their effects on resource utilization, performance, and overall cloud expenditure.

V. SCOPE OF THE STUDY

“Mostly operational costs arise at the infrastructure as a Service and Platform as a Service layers. Hence, this research is about optimizing the cloud cost of these service layers”.

The scope of optimization strategies includes autoscaling, right-sizing, spot and reserved instances, storage lifecycle policies, serverless computing, and cost monitoring.

It would study cloud usage data, apply optimization techniques, and then look at the results in terms of cost savings and performance metrics.

There are two major exceptions. Detailed cost models of hybrid cloud billing and any vendor-specific proprietary optimization besides standard procedures are out of our domain.

“We will implement the groundwork using commonly available tools and services of typical cloud service providers such as AWS and GCP”.

Cloud Cost Optimization Techniques

1. "Right Sizing one of the major cost reductions in computing resources techniques is right sizing. It aligns the compute resources allocated with the actual requirements of the workload.

By determining the right requirements for CPU, memory and storage. Companies can reduce their wastefulness and save excess charges. Costing $0.388472 per hour.

GCP's n2 standard 8 instances would cost $0.776944 per hour for n2 standard 16 instances in GCP.

If a workload were to need only 8 vCPUs, then right sizing to a n2 standard 8 instances from an over-provisioned n2-standard-16 instance would save $0.388472 per hour. Over time, for long-running workloads, those savings can add up. Optimization of compute costs may be achieved through right-sizing compute resources, resource monitoring, analysing performance metrics, or by taking advantage of tools offered by the cloud provider and those available from third-party providers".

2. "Autoscaling is every dynamic form of resource allocation that changes compute resources based on workload demand.

By scaling resources dynamically upwards or downwards, enterprises can keep compute capacity in alignment with application needs, which are perpetually changing. Autoscaling ensures that there is proper utilization of resources with no overprovisioning during low demand periods and increased response time during heavy loads. Enterprises are billed for compute resources applied for hence a great cost benefit".

3. "Spot Instances are considered a cheap choice for non-critical works. Cloud providers sell spare compute capacity at heavy discounts so that enterprises can bid for those instances.

These spots give a considerable amount of savings compared to on-demand or reserved instances. However, keep in mind that the spot instances can be interrupted when the spot price exceeds the bid price. Hence, they have to be fault-tolerant, flexible workloads that can be interrupted.

But, because Spot Instances have no 100% guarantee of availability, in the event an instance is lost, users get notified of about two minutes warning to manually terminate the instance's applications. One way to minimize the chance of such interruptions is the use of Spot Fleets.

A Spot Fleet is a collection of spot instances that arrive as one unit".

4. Reserved Instances Reserved Instances allow enterprises to enter into a discounted pricing scheme by agreeing for using certain compute capacity for a given time. Paying upfront for reservation leads to a reduced hourly charge compared to paying hourly for the demand. Generally, reserved instances should operate under workloads in whose requirements exhibit predictable and constant demand.

5. Storage- Strategies and practices that can be used to employ cost-optimization for storage in a cloud environment are discussed in this section. Proper utilization of storage resources and cost-saving methods on companies' part can help them to reduce storage costs while at the same time making data accessible and reliable.

6. Data deduplication and compression are the key to cost optimization in cloud-storage environments. "Since the older ones attempt to remove redundant or never-to-be-used data, organizations may drastically reduce storage requirements and the corresponding costs. Deduplication identifies similar data segments and deletes them, whereas compression caters to size reduction by data encoding via its more efficient coding algorithm. In tandem, these techniques allow significant cost advantages in terms of reducing the physical media needed for storage and succumbing to the pressure from data growth. The second most important advantage of data deduplication is the eliminating of redundant data copies. This kind of scenario usually occurs within companies when multiple users or applications store duplicate or similar files. Deduplication brings about savings by reducing storage capacity through identifying and keeping only unique data segments. Besides that, compression techniques increase storage efficiency by reducing file or data block sizes.

Besides saving on storage capacity, deduplication and compression lead to other cost savings, such as down the road during data transfer.
Significantly reducing the storage footprint saves organizations money while transferring data between cloud storage tiers or between regions. Lesser data sizes enhance transfer speeds, giving another layer of PD enhancement.
Deduplication and compression could also enhance backup and disaster recovery cameras and reduced cost for application downtime. Disaster downtime and costs get reduced with deduplication and compression efforts".
Smaller data volumes mean a reduced time required for backup or recovery actions. Implementing data deduplication and/or compression requires knowledge of data access patterns, application requirements, and computation overheads.
Choosing deduplication and compression algorithms that provide reasonable storage savings yet account for the cost in processing time and data access times will be key.
Points to consider would be evaluating trade-offs between storage cost saving and computation resource requirements.

VI. CASE STUDIES

Cost optimization in the real world should be applied wherever businesses intend to streamline operations and reduce their infrastructural expenses.

The case studies of famous businesses included Amazon and Pinterest-are areas where we'd like to present and discuss various occasions in which cost-optimization strategies have been used effectively.

"Bearing in mind the ample information available about the implementation of the various techniques discussed in this paper, we intend to learn valuable lessons and best practices for efficient cost optimization by studying how these organizations were able to address their infrastructure challenges and realize huge cost savings".

Through these real-world cases, we hope to illustrate just how diverse the strategies businesses employ to cut infrastructure costs are, especially in terms of maintaining or improving performance and scalability.

1. Amazon Prime Video

"Prime Video is a very rapidly growing streaming platform with a wide choice of movies and TV shows". There are three components under discussion here related to how Prime Video's cost-cutting methods re-architected this system for lesser infrastructure and audio-video monitoring service costs.

A. "Background The audio-video monitoring feature on Prime Video was developed using a distributed microservice architecture, which was a type of architecture that contained independent services, each responsible for monitoring something about audio or video quality".
For example, there would be a service to monitor the audio loudness and a separate service to monitor the video bitrate. The microservice architecture provided benefits such as easy development, deployment, and scaling of a new service by independently running more instances of the service. However, the microservice architecture has disadvantages including difficulty controlling the service maintenance aspect of all the independent services, and the other disadvantage is the independent scaling not being very successfully scalable.
"Hence, to solve these challenges, Prime Video went ahead to define the audio-video monitoring service as a monolith. A monolith is one single centralized service that manages all levels of audio or video quality monitoring. By merging all services into one monolith, it increased the ease of scaling and maintaining the audio-video monitoring service, since it was no longer necessary to copy data between the services—thus making the service more efficient.

B. Cost Optimization Techniques "The key to cost control Any AWS Step Functions is a serverless orchestration service for multiple AWS services to coordinate execution. The tool is powerful but also expensive. The audio-video monitoring system in Prime Video somehow orchestrated the data flow through that service using AWS Step Functions. This actually turned into a bottleneck.
Every single second of the stream transitioning through multiple different states. So, account-limit on

Prime Video was reached very quickly through-limit. Since AWS Step Functions also charge per state transition, the whole cost of all of its building blocks was too high for the solution to be taken up more widely. Furthermore, Prime Video's audio-video monitoring service was giving the solution to store the video frames in Amazon S3, and this service made tons of Tier-1 requests to Amazon S3. Being the most expensive calls possible, Tier-1 calls are the least desirable calls one can make to Amazon S3. Hence, Prime Video is paying a hefty amount to Amazon S3 for storing video frames. On the other hand, the number of video frames that need to be stored may vary with traffic volume.

Prime Video placed an audio-video monitoring service under redesigning, turning it into monolith application owing to bottlenecks and cost issues. All components started running in a single process, and hence, AWS Step Functions and Amazon S3 went out of the picture. With this pricing idea removed, Prime Video saved substantially from the state transitions and Tier-1 calls. The monolith architecture also scaled up and down as the need arose. This was important, as an increase in traffic was expected on the service. Finally, shifting the solution to AWS EC2 and AWS ECS proved useful, as it enabled Prime Video to avail the long-term compute savings plans from AWS EC2 along with EC2 features like autoscaling with respect to traffic load, further helping reduce costs.

Results Changes reduced the running cost of its audio-video monitoring service by 90%. And it was able to handle much higher traffic".

C. Lessons: "Some of the primary lessons to be learned from this case study of Amazon Prime Video:

- It becomes costly and difficult to scale distributed microservices architectures.
- Monolithic architectures do have the potential to provide cost savings and scalability.
- You can use compute saving plans to save money for Amazon EC2 usage.
- Spot instances, reserved instances, and autoscaling can also be another potential solution for you to save your Amazon EC2 costs".

VII. INDUSTRY EXAMPLE

1. Background- The transformation was taking place within a mid-sized IT services company because of the ever-increasing AWS storage costs. As per a cost-containing evaluation, EBS snapshot charges constituted a major portion of the expenses. These snapshots are created during a variety of activities including testing, staging, and production; however, once the corresponding EC2 instances get terminated or the project gets closed, the snapshots still linger on in the account. Since AWS levies charge of $0.05 for storage of snapshots per GB-month, the accumulation of such unused or stale snapshots became recurring unnoticed expenses that add considerably to the huge cloud bill.

2. **Technique-** The DevOps team provided a solution by setting up and deploying an AWS Lambda function that automates the cleanup of stale EBS snapshots. The function first fetches all snapshots owned by the account; then it obtains the list of running EC2 instances and cross-references one by one with each snapshot to ascertain whether the snapshot's parent volume has been attached to any instance, either running or stopped. Those snapshots whose parent volumes are not associated with any active instance are considered stale and get deleted. To give it longevity, the Lambda function is scheduled to run periodically via CloudWatch Events from Amazon to continuously monitor and clean up unused snapshots without needing manual intervention".

3. **Result** For the execution cycle one, thousands of stale snapshots were identified and deleted by a lambda function for a total capacity of nearly 3 TB, saving close to $150 per month, the difference equating to roughly $1,800 per year. Cost savings were, however, not the only area of impact; the operational efficiency and cloud governance team at the company also improved since the retention of data was ceased. This is a good example of how targeted automation in cloud environments can yield long-term benefits related to cost optimization and resource management.

VIII. FUTURE SCOPE AND LIMITATIONS

1. Future Scope:

This presents numerous opportunities for further research and actual implementation in the domain of cloud cost optimization. While this work proposes techniques for automating the identification and deletion of stale EBS snapshots, in principle, the same methods could apply to other costly resources like idle EC2 instances, orphaned EBS volumes, underutilized load balancers, and unused Elastic IPs.

Predictive analytics using machine learning models could be further incorporated to transform cost management into a more proactive exercise, wherein predictive usage planning and right-sizing suggestions preempt potential cost-waste.

Another research avenue worth exploring would be to integrate FinOps practices into this automation framework to cast financial accountability side-by-side with technical efficiency. In a similar vein, evolution of the solution into a multi-cloud one would allow a universally recognized cost-optimization layer over an organization's hybrid or multi-cloud stack. Lastly, a real-time visualization and alerting dashboard would provide decision makers a step closer to actionable insights to govern against.

2. Limitations:

Yet, with all the successes, the project does not come without its set of constraints. First, this solution is AWS-specific and targets EBS snapshots; thus, its direct application to other cloud providers is diminished unless adapted for use. Second, the Lambda function assumes that all snapshots not associated with volumes in use are unwanted; however, in certain scenarios, organizations may have compliance or auditing requirements specifying longer retention periods for snapshots, which the automatic deletion process may not honour unless it is explicitly configured to do so. Third, the cost savings resulting from this method will depend on the volume of unused resources, whereby organizations running smaller workloads may not accrue any significant savings. Lastly, any automation will reduce human intervention but increase dependency on IAM permissions and correctly configured policies that, if neglected, could lead to accidental data loss or a security compromise.

IX. CONCLUSIONS

The project successfully showcased cloud cost optimization via smart automation fixating on the identification and deletion of stale EBS snapshots using AWS Lambda. By identifying the hidden storage costs through a simplified automation function, which it then perpetually schedules via CloudWatch Events, it outlined a practical framework for minimizing unneeded expenditures in cloud environments. The implementation yielded exceptionally good results, removing close to 3 TB of snapshots sitting idle, saving roughly US $150 a month or US $1,800 annually on a direct basis. The solution offers more than just savings; it also offered improved operational efficiencies, governance, and utilization of assets with little need for human interaction filed under managing lifecycle of cloud assets.

This project demonstrates the fact that tiny and strategic automation efforts can provide major returns in the long run when applied to cloud-costing efficiency. In slight variations, it can provide a reusable model to extend similar practice across other cloud resources for sustainable financial savings with stronger alignment to business objectives from cloud usage.

X. REFERENCES

1. Bidikov, V., Gusev, M., Markozanov, V.: Network traffic impact on cloud usage at different providers. In: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO). pp. 847–852 (2022)

2. Sun, X., Zhuo, X., Wang, Z.: A survey of pricing aware traffic engineering in cloud computing. Journal of Internet Technology 21, 357–364 (2020)

3. Lee, G., Lin, J., Liu, C., Lorek, A., Ryaboy, D.: The unified log-ging infrastructure for data analytics at twitter. ArXiv (2012), https://doi.org/10.48550/arxiv.1208.4171

4. Collet, Y., Kucherawy, M.: Zstandard

compression and the "application/zstd" media type. https://www.rfc-editor.org/rfc/rfc8878 (2019), last accessed on 2023/07/22

5. Handte, F., Collet, Y., Terrell, N.: Zstandard: How facebook increased compression speed. https://engineering.fb.com/2018/12/19/core-data/zstandard/ (2018)

6. Microsoft: Database Advisor performance recommendations for Azure SQL Database. https://learn.microsoft.com/en-us/azure/azure-sql/database/database- advisor-implement-performance-recommendations?view=azuresql-db, last ac- cessed on 2023/07/22

7. Google Cloud Platform (GCP): Gcp active assist recommenders. https://cloud.google.com/recommender/docs/recommenders (2023), last ac- cessed on 2023/07/22

8. Amazon Web Services (AWS): Amazon Virtual Private Cloud IP Address Manager (IPAM). https://docs.aws.amazon.com/vpc/latest/ipam/what-it-