Deep Science Publishing
https://doi.org/10.70593/978-81-984306-5-6_7

Chapter 7

Cost optimization in hybrid cloud environments

Ravi Kumar Vankayalapati

Cloud AI ML Engineer, Equinix, Dallas, USA.
ravikumar.vankayalapti.research@gmail.com

Abstract

Cost optimization in hybrid cloud environments is a critical consideration for organizations seeking to balance the flexibility of public clouds with the control of private clouds. By strategically managing workloads across both environments, businesses can maximize resource efficiency and minimize unnecessary expenditures. This abstract explores key strategies for cost optimization, including workload placement, right-sizing of resources, and the use of auto-scaling features in public clouds. It also highlights the importance of monitoring tools, cost management platforms, and financial governance models to track and control cloud spending. By implementing these approaches, organizations can ensure a cost-effective hybrid cloud strategy while maintaining performance and scalability.

Keywords

Cost Optimization, Hybrid Cloud, Cloud Cost Management, Workload Placement, Resource Efficiency, Auto-Scaling, Public Cloud, Private Cloud, Cloud Strategy, Cost Management Tools, Financial Governance, Cloud Spending, Cloud Scalability, Hybrid Cloud Strategy, Cloud Resource Right-Sizing, Performance Optimization, Cloud Monitoring.

7.1. Introduction

Cloud computing has revolutionized the way organizations manage their IT infrastructure. The advent of cloud services completely changed the view of computing and almost all IT services because of the pay-as-you-go model. However, with this came a significant question for IT managers who are supposed to optimize the cloud budget allocated to them. The goal is not just enormous scale, elasticity, and redundancy; it’s also about doing all of that as cost-effectively as possible. Recent adoption of hybrid cloud in the industry adds a significant layer of complexity in the budgeting task as on-prem servers and cloud services are not in the same monitoring or billing domain (Rama, 2022). Owing to multiple cloud strata with inter-cloud communication, the problem statement is extended and defined as a problem of allocating resources in a hybrid-extensive cloud to multiple applications in a combination of on-prem and cloud infrastructures. To tackle this dilemma, an in-depth plan to cross cloud and on-prem structures with high monitoring is carried out. This plan uses an off-the-shelf monitoring set of tools for detailed metering and profiling of resources. From the collected data, a mathematical model is formed and the problem described above is reformulated as a complex combinatorial optimization that is nontrivial to solve precisely even for relatively small problem instances. In this light, the question is optimized using a profoundly scalable greedy algorithm based on the concept of maximum visibility. Cloud computing has fundamentally transformed how organizations manage IT infrastructure, offering scalability, flexibility, and cost-efficiency through a pay-as-you-go model. However, this shift brings challenges for IT managers who must optimize cloud spending while ensuring that resources are used efficiently. The complexity intensifies with the growing adoption of hybrid cloud environments, where on-premise servers and cloud services operate in distinct monitoring and billing domains (Danda, 2024). This necessitates a sophisticated approach to resource allocation across both on-prem and cloud infrastructures. To address this challenge, an integrated plan is developed, utilizing a suite of off-the-shelf monitoring tools to meticulously track and profile resource usage. The collected data is then used to build a mathematical model, transforming the problem into a complex combinatorial optimization issue. Due to its complexity, solving this problem exactly is difficult, even for small instances. As a result, a scalable greedy algorithm is employed, leveraging the concept of maximum visibility to efficiently allocate resources across the hybrid cloud, ensuring cost-effective optimization while maintaining performance and scalability.

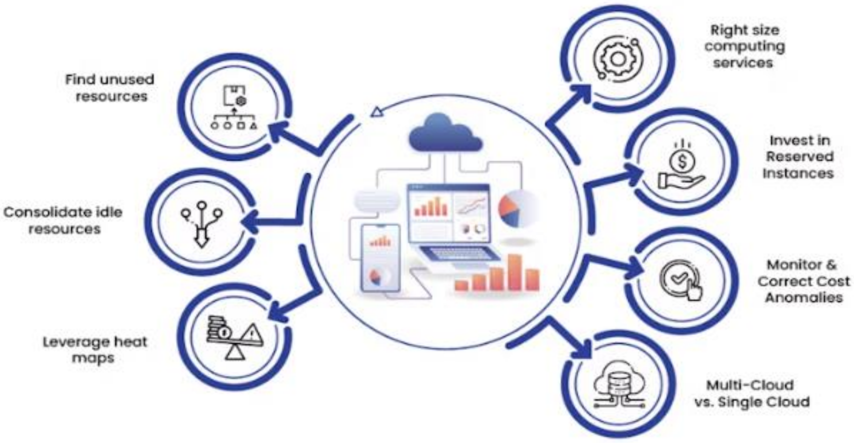

Find unused resources
Consolidate idle resources
Leverage heat maps
Right size computing services
Invest in Reserved Instances
Monitor & Correct Cost Anomalies
Multi-Cloud vs. Single Cloud

Fig 7.1: Cloud Cost Optimization

7.2. Understanding Hybrid Cloud Environments

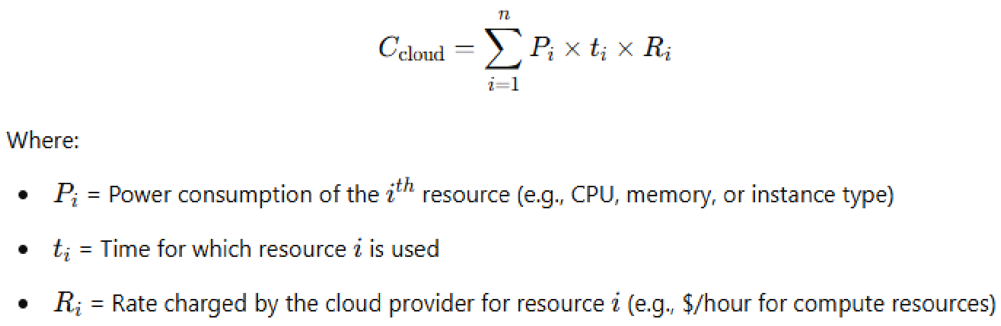
This section explains the environment of the hybrid cloud and the involved pricing structures, particularly for a hybrid cloud with multiple providers. In the last decade, cloud computing has received a lot of attention due to the benefits it offers, such as flexibility, scalability, pay-per-use billing, and many more qualities that help organizations manage their IT functions. One of the most important concepts in cloud computing is hybrid cloud. A hybrid cloud refers to the use of many cloud services to support a single application and the composition of these services can vary according to the desired service quality. For many applications deployed on a hybrid cloud (HC), there is a need for usage of multiple cloud service providers (CSPs). Therefore, pricing rules offered by them contribute to a high cost of the application owner. A good knowledge of pricing rules is very important for making accurate decisions about resources, tasks or services that will be used in HC. However, most of the pricing structures are complex and sometimes difficult to understand for users (Syed, 2022). This complexity can be attributed to the large number of pricing rules and the presence of many parameters, so many researchers have tried to escalate this problem.

There is much research on cloud pricing models, but to the best of the knowledge, little information was published on formats for pricing structures to address the needs of service-level planning and BMC. So, the purpose of this section is to present detailed pricing structures of those CSPs that are commonly used in the HC with a single cloud provider (2C1P) and many cloud providers (2CnP).

7.2.1. Definition and Components

Cloud computing has revolutionized the way organizations manage their IT infrastructure. Management of the IT infrastructure cost is one of the most important issues for any organization as it supports business activities. With the advent of cloud computing, world organizations have started migrating their IT infrastructure to the cloud due to its benefits such as agility, scalability, and reduced capital expenditure. However, the cloud introduces many new challenges. In a cloud environment, infrastructure, and services are provided over the internet on a pay-as-you-go basis. As a result, the operational expense (OpEx) increases in a sporadic fashion that is not as transparent as the traditional capital expenditure. So, managing the cloud cost becomes more critical. If the costs become out of control it becomes higher than the benefit derived from the cloud. This type of cost is known as the Cloud Bill Shock problem. Hence, Cloud Cost Optimization is required. A Cloud computing environment consists of several components that are required for managing different tasks. The following should be the necessary components that should be included in the Cloud Cost Optimization framework like Cloud Consumer, Cloud Provider, Resource, Applications, Services, Metering, Cloud Broker, Orchestration (Service Integration), Policy (Governance), Billing & Rating, and Optimize resources based on current performance and future requirements.

Equation 1: Cloud Cost Estimation with Dynamic Scaling



$$C_{\text{cloud}} = \sum_{i=1}^{n} P_i \times t_i \times R_i$$

Where:

- P_i = Power consumption of the i^{th} resource (e.g., CPU, memory, or instance type)
- t_i = Time for which resource i is used
- R_i = Rate charged by the cloud provider for resource i (e.g., $/hour for compute resources)

7.2.2. Benefits and Challenges

Cloud computing has revolutionized the way organizations manage many IT needs by providing an innovative technology that is scalable, elastic, and capable of providing access to a vast pool of data storage and resources. Emerging companies, corporate industries, governmental organizations, academic institutes, and a wide variety of companies are taking advantage of cloud computing to efficiently handle their IT infrastructure (Nampalli, 2021). Despite its varied advantages, the fact that it allows users to access storage and on-demand information from anywhere is viewed as the most advantageous feature of cloud computing.

Cloud computing has introduced a wide range of challenges that have accumulated over time, one of the most pressing of which is cloud costs. Organizations, on the other hand, must pay for the storage and services they utilize. The most common billing procedure is to charge by the number of resources being used. As a result, to minimize costs, businesses such as colleges, e-commerce, tech firms, and governmental companies need to think about resource use. Companies that do not properly manage cloud costs end up spending millions of dollars on cloud services. On another perspective, the cloud pricing and payment patterns are very complicated and it is hard to accurately measure, predict, and compare costs. Local machine prices, time-based estimation, cross supplier cost evaluation, expense trend forecasting, and cost allocation are just a few of the challenges. Nonetheless, it is possible to approach a wide collection of data from a variety of sources and make proper use of them to manage costs. Cloud pricing, cost monitoring, and policy-driven resource allocation are strategies that can be used for this. Organizations can employ a number of mechanisms to minimize costs while still providing cloud services and satisfying service-level agreements. Appropriate expense management analysis could help organizations save over a million dollars a year in cloud costs.

7.3. Cost Factors in Hybrid Clouds

Cost optimization in cloud environments is the process of adjusting costs to maximize the value of the service provided while retaining the same level of function. Optimal cloud costs should support application performance, ensure a certain level of security, and minimize costs for both service providers and users. If a system uses an excessive amount of resources, it is over-provisioned. In contrast, under-provisioned systems are incapable of transferring the requested task or have low performance. Over-provisioning and underprovisioning are some of the most common resource allocation problems in cloud environments and critical optimization problems.

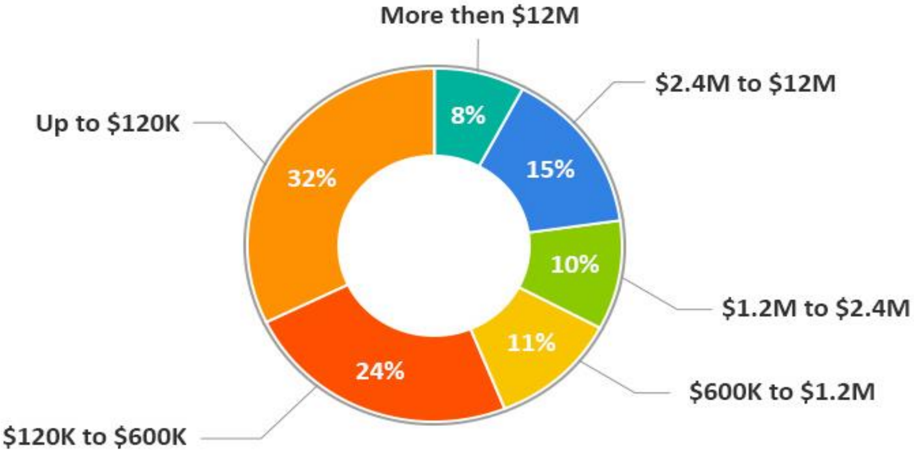


Fig: Cloud Cost Optimization and Management: Trends, Analysis and Strategy

Cloud computing has revolutionized the management of IT infrastructure in different organizations. However, besides the many benefits of cloud services, cloud environments have evolved new problems for organizational users (Danda, 2024). One of these significant challenges is cloud costs, particularly the costs of profit-oriented service providers. Organizations have difficulties in determining the organization's cost-effective resources in the cloud, and strategically measuring the resulting effect may be impossible. Additionally, unexpected costs may arise from other cloud services. Due to the cost pressure, organizations have begun to trade off the performance of cloud systems, which is usually degraded. From the cloud service provider's viewpoint, managing the client's application QoS and service costs is a critical problem. Commonly, resource allocation processes are used for managing the resources that data centers need to host. This issue involves partitioning the resources among client applications adapting with changes in workload demands which usually come in the form of application workload changes, technological or hardware upgrades, or outages. The decision constraints are principally directed to service level agreements (SLAs). In case of SLA violations the client gets some limit of the punishment.

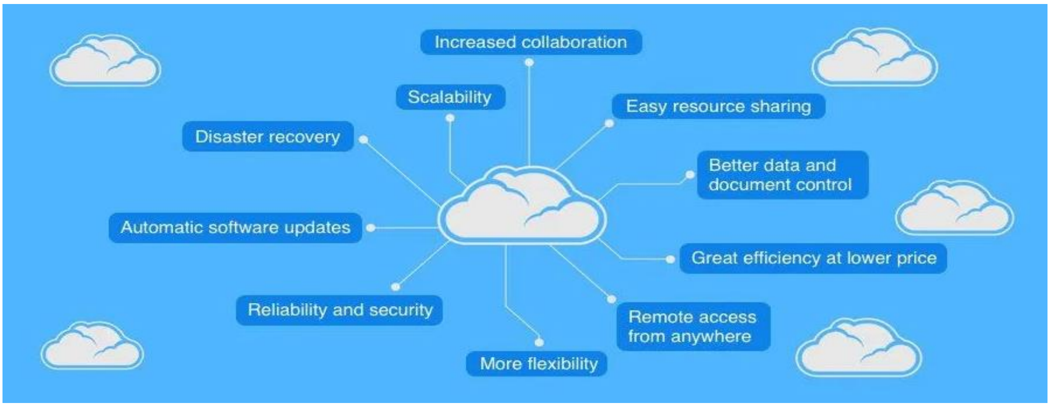
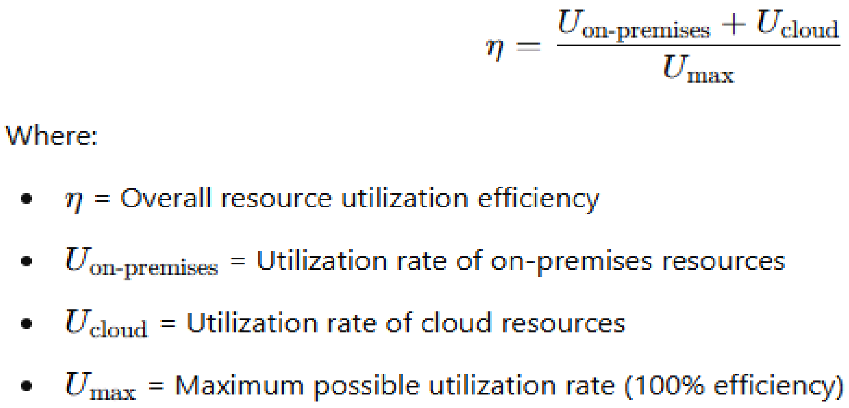

Increased collaboration
Scalability
Easy resource sharing
Disaster recovery
Better data and document control
Automatic software updates
Great efficiency at lower price
Reliability and security
Remote access from anywhere
More flexibility

Fig 7.2: Cost Factors in Hybrid Clouds

7.4. Cost Optimization Strategies

Comprehensive exploration of cloud cost optimization techniques: As businesses increasingly move to the cloud, they often find that traditional methods for managing IT costs do not directly map to cloud pricing models. This has vaulted the topic of cloud cost optimization into prominence in recent years (Kothapalli et al., 2022). A methodical approach that examines a comprehensive exploration of various cloud cost optimization techniques is presented. Cloud cost optimization encompasses a variety of strategies that can help enterprises effectively manage and optimize cloud expenses while maintaining performance, reliability, and security. By understanding and implementing these techniques, organizations can gain better control over their cloud costs and maximize the value of cloud investments. From financial engineering and cost expectations to real-time management and resource scheduling, this process aims to work towards a broad, yet systematic, view on how to optimize cloud costs. This manageable approach examines practical strategies that enable enterprises to achieve cost efficiency and financial optimization in cloud environments. These strategies include resource allocation, workload optimization, automation, governance, network infrastructure, and IaaS/PaaS/SaaS platforms. In terms of workload optimization, a common technique is containerization, which packages applications and their dependencies in isolated environments (Subhash et al., 2022). Containers provide a lightweight and portable application runtime, leading to reduced resource overhead and improved system efficiency. Enterprises can leverage these benefits to maximize resource utilization, scale applications appropriately, and optimize for cloud spend. Efforts toward automation cover various facets of cloud management life cycle, ranging from monitoring and analysis to decision-making and policy enforcement. Automation through machine learning, predictive modeling, and autoscaling technologies can help enterprises react quickly to workload changes and ensure optimal system performance.

Equation 2: Resource Utilization Efficiency in Hybrid Cloud



$$\eta = \frac{U_{\text{on-premises}} + U_{\text{cloud}}}{U_{\max}}$$

Where:

- η = Overall resource utilization efficiency
- $U_{\text{on-premises}}$ = Utilization rate of on-premises resources
- U_{cloud} = Utilization rate of cloud resources
- $U_{\max}$ = Maximum possible utilization rate (100% efficiency)

7.4.1. Resource Sizing and Allocation

The actual challenge is to define a strategy that is able to place workloads each hour considering short-term SLA obligations, and to take resizing decisions on virtual machines when needs change. Different strategies can be thought, depending on the horizon of scheduling and resizing activity, but a general approach is here reported and applied to the business case: a set of well-known and easy computable heuristics is proposed, as this problem is quite novel and hard to manage, along with an optimal approach that would become handy when the set of clusters eligible for each application role is limited. To assess the significance of the gap with respect to the optimal approach, a lower bound is obtained by performing offline scheduling and resizing decisions on the jobs collected from history. Further, the experiments consider that low complexity strategies can be supported by computational tools, and need the effectiveness of the lower bound to better describe the quality of forecast and optimization problems.

7.4.2. Reserved Instances and Savings Plans

The use of on-demand instances hosted on the borrowing public cloud causes most of the cloud cost. Reserved Instances and Savings Plans, a pricing model that offers a discount for committing to a particular level of resource usage, are money-saving strategies. Reserved Instances were introduced in 2009 making it the first public cloud provider to offer a way to reduce computing costs. Since then, other public cloud providers have introduced similar features. Reserving instances in exchange for a usage commitment of between 1 to 3 years often results in significant discounts compared to using them on-demand.

Using On-Demand instances when the load is regular or can be predicted with a reasonable margin is an inefficient way. There is an opportunity of purchasing a specific hardware configuration in an instance, and also purchasing the time for using it (Sondinti et al., 2023). These reservations can be used for a 12-month or 3-year period and for an operating system of the users' choice. Another provider offers a similar reservation model, when purchasing in 1 and 3 years a specific hardware configuration in an instance, a discount is obtained. A similar offer using a committed use discount also offers the opportunity to buy hardware configuration for a certain period of time.

Savings Plans is a relatively new pricing model introduced in 2019 and then adopted by another provider as Committed Use Discounts and by another as Reserved VM Instances. If a business uses some specific usage commitments model, Savings Plans provide greater flexibility and savings compared to traditional Reserved Instances. Unlike traditional Reserved Instances, with Savings Plans, a business commits to a certain amount of usage in exchange for a discounted rate on their bill, this is much easier than trying to estimate the specific instance types the business will need in the future. Almost all Substitute cloud providers have adopted a business model based on typical cloud providers.

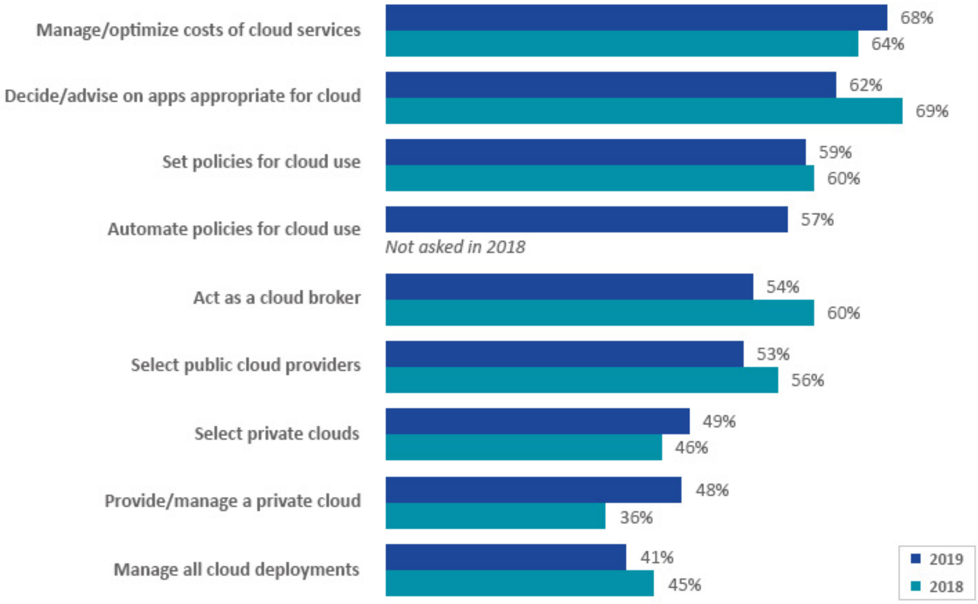
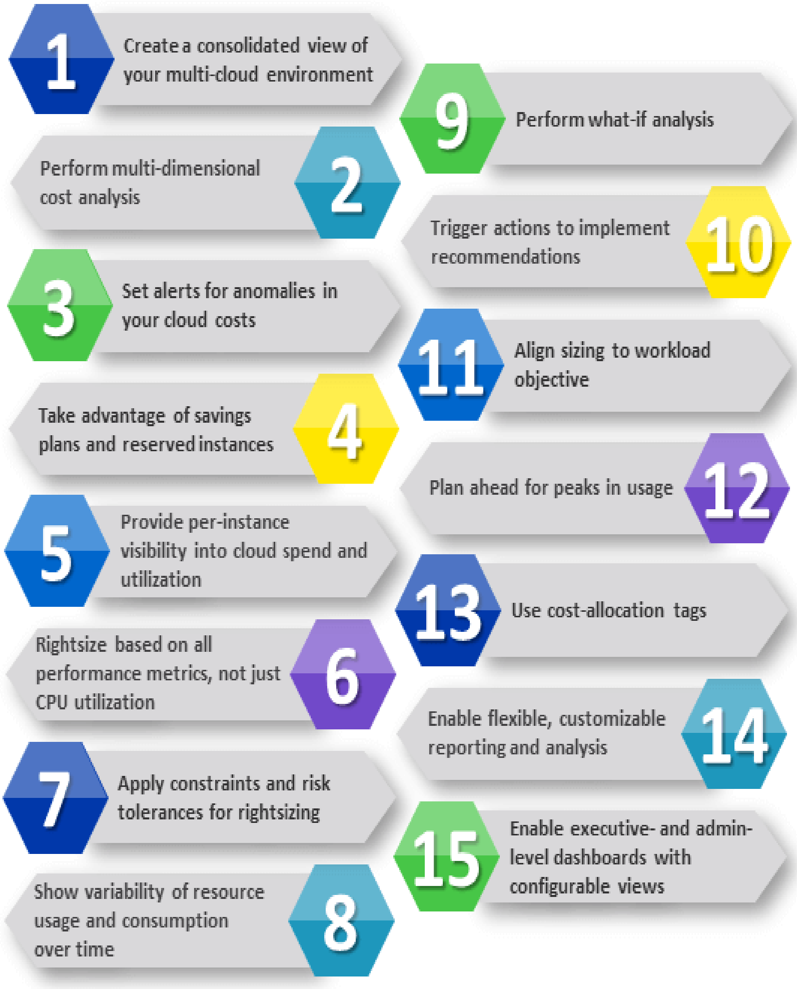


Fig: 2019 State of the Cloud Report on 'Cloud Spending

7.5. Case Studies and Best Practices

As cloud services have become the norm for the majority of IT applications, it is critical to ensure that an organization's cloud system is budget-friendly and uses resources effectively. It is essential not only to achieve an overall aim of calculating, measuring, and increasing IT service quality and efficiency, but also to manage the financial investment in cloud technologies and services. As a result, optimization costs for the cloud should match performance tuning activities, load prediction, and capacity management to recognize possibilities to enhance usage and save money (Vankayalapati et al., 2023). This research explores pricing and cost analysis on current cloud services to select appropriate types and data centers for application deployment. This is pursued by strategies oriented to effective use of cloud resources so that applications and workloads provided by the hybrid cloud deployment model can be optimized.

Based on the findings, the a priori approach can be implemented along with reactive techniques. A priori approaches should be mostly used on cost saving techniques for IaaS services and on their initial deployment to determine how many those services should be deployed additionally or removed to it. In contrast, cost analysis and reactive optimization approaches should take time regularly after services are utilized, modified, or removed. Part of the best practices will also be based on adapting application architectures. Adding a queue layer and moving components to data service or storage machines as much as possible can significantly increase both throughput and quota consumption. Because of cost inefficiencies, homogeneous scaling of application workers, or inappropriate placement of service components, many applications won't be able to use the benefits of the public cloud. On the other hand, efficient, low-cost applications benefit most from adhering to these best practices.



1 Create a consolidated view of your multi-cloud environment
2 Perform multi-dimensional cost analysis
3 Set alerts for anomalies in your cloud costs
4 Take advantage of savings plans and reserved instances
5 Provide per-instance visibility into cloud spend and utilization
6 Rightsize based on all performance metrics, not just CPU utilization
7 Apply constraints and risk tolerances for rightsizing
8 Show variability of resource usage and consumption over time
9 Perform what-if analysis
10 Trigger actions to implement recommendations
11 Align sizing to workload objective
12 Plan ahead for peaks in usage
13 Use cost-allocation tags
14 Enable flexible, customizable reporting and analysis
15 Enable executive- and admin-level dashboards with configurable views

Fig 7.3: Cloud Cost Management Best Practices

7.6. Conclusion

Since Elastic Compute Cloud (EC2) was introduced, the applications of cloud computing have been revolutionized. Cloud computing offers utility (on-demand) based access to shared computing resources. Organizations no longer need to maintain their own physical IT infrastructure. They can now rent the required computing, storage and networking through services provided by public cloud vendors. Until today, the cloud vendors have vastly improved the services, resources and the billing system. The cloud computing platforms also provision many complex systems, services and technologies.

Generally, the upfront cost of cloud computing is less, but over a period of time costs increase because of various reasons (Maguluri et al., 2022). First of all, organizations do not recognize the cloud service and resources which they actually need and eventually the cost of actively running cloud resources increases. Also, they don't know the proper configuration and tuning of resources they are active with which leads to increased cost. In Research and Development environments, it is hard to predict the exact requirement, in this case, resources keep running (though not used) for development and testing purposes. There are also some hidden costs like the data transfer rate between regions is high. At different regions using separate database services or transferring data between them leads to extra billed cost. Similarly, if data transfer to the internet is high due to heavy API calls it will result in increased cost.

Similarly, consumers fail to stop the resources when not in use. With the flexibility of cloud resources and services just one API can terminate all the running resources even with tags. The environments can be shaved according to the closure of business hours programmatically even on hard software cache hits. So there are some complexities and requirements to make the cloud worthy of the organizations. This paper deeply investigates the cloud services and resources and concludes with some strategies to make the cloud resource and services to the fullest use of the organizations.

7.6.1. Future Trends

As the widespread adoption of cloud-based infrastructures continues to bring ever-greater scale and complexity to online services, networked systems developers are forced to take into account the financial efficacy of their designs and proposed systems. Cloud-based online services typically adopt some form of fixed capacity provisioning as an initial step to building a new service. Many public cloud providers commit new virtual machines and network, and/or storage fabric on a “pay as you go” basis. The provider is responsible for the physical implementation of brought services and the consumer concerned with the presence and performance of the hired service. Interestingly, networked systems developers often observe that a small portion of finished services proves enduringly more popular than the majority. This continuing popularity is associated with a wide range of psychological and technological factors but can manifest in long periods where multifaceted software services are established and running on heavyweight cloud provisioning but at resource utilisation rates that average less than 15% of capacity. Such idleness typically coincides with the presence of an external event such as peak advertising, education term, sporting fixtures, etc., that renders it financially impractical to further reduce service fixed costs for the period of the event. In the light of improving service design practice, a theoretical model is presented that attempts to guide the software service developer on selecting the optimal step size of cloud-procured network and server resources given the projected step-change in demand for the service. The model takes as input the level of service popularity, the fixed cost of procuring cloud resources, the demand profile for the service during its career run-time, and the lifetime of the service. The model predicts economically optimal resource steps given a meta-model representing the cloud provider's pricing schedule. Relying on the step size of cloud resources notified by the model, it is shown that an idealized online service could achieve a nearly 33% reduction in fixed costs over a six-month billing window compared to a naive step-change in resource procurement of the same services. Plausibility of the model is assessed by empirical validation with two toy and real-world services from the public cloud environment. Further discussion of the proposed model considers practical aspects of its limitations, exploitation, and possible implementation.

References

Danda, R. R. (2024). The Role of Machine Learning Algorithms in Enhancing Wellness Programs and Reducing Healthcare Costs. Utilitas Mathematica, 121, 352-364.

Danda, R. R. (2024). Using AI-Powered Analysis for Optimizing Prescription Drug Plans among Seniors: Trends and Future Directions. Nanotechnology Perceptions, 2644-2661.

Kothapalli Sondinti, L. R., & Yasmeen, Z. (2022). Analyzing Behavioral Trends in Credit Card Fraud Patterns: Leveraging Federated Learning and Privacy-Preserving Artificial Intelligence Frameworks. Universal Journal of Business and Management, 2(1), 1224. Retrieved from https://www.scipublications.com/journal/index.php/ujbm/article/view/1224

Maguluri, K. K., Pandugula, C., Kalisetty, S., & Mallesham, G. (2022). Advancing Pain Medicine with AI and Neural Networks: Predictive Analytics and Personalized Treatment Plans for Chronic and Acute Pain Managements. In Journal of Artificial Intelligence and Big Data (Vol.

2, Issue 1, pp. 112–126). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2022.1201

Nampalli, R. C. R. (2021). Leveraging AI in Urban Traffic Management: Addressing Congestion and Traffic Flow with Intelligent Systems. In Journal of Artificial Intelligence and Big Data (Vol. 1, Issue 1, pp. 86–99). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2021.1151

Rama Chandra Rao Nampalli. (2022). Deep Learning-Based Predictive Models For Rail Signaling And Control Systems: Improving Operational Efficiency And Safety. Migration Letters, 19(6), 1065–1077. Retrieved from https://migrationletters.com/index.php/ml/article/view/11335

Sondinti, L. R. K., Kalisetty, S., Polineni, T. N. S., & abhireddy, N. (2023). Towards Quantum-Enhanced Cloud Platforms: Bridging Classical and Quantum Computing for Future Workloads. In Journal for ReAttach Therapy and Developmental Diversities. Green Publication. https://doi.org/10.53555/jrtdd.v6i10s(2).3347

Subhash Polineni, T. N., Pandugula, C., & Azith Teja Ganti, V. K. (2022). AI-Driven Automation in Monitoring Post-Operative Complications Across Health Systems. Global Journal of Medical Case Reports, 2(1), 1225. Retrieved from https://www.scipublications.com/journal/index.php/gjmcr/article/view/1225

Syed, S. (2022). Integrating Predictive Analytics Into Manufacturing Finance: A Case Study On Cost Control And Zero-Carbon Goals In Automotive Production. Migration Letters, 19(6), 1078-1090.

Vankayalapati, R. K., Sondinti, L. R., Kalisetty, S., & Valiki, S. (2023). Unifying Edge and Cloud Computing: A Framework for Distributed AI and Real-Time Processing. In Journal for ReAttach Therapy and Developmental Diversities. Green Publication. https://doi.org/10.53555/jrtdd.v6i9s(2).3348