ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Comprehensive Review on Optimizing Resource Allocation in Cloud Computing for Cost Efficiency

Ankur Mahida

Subject Matter Expert (SME), Barclays, USA

ABSTRACT

The cloud computing model has become very popular among organizations as they search for the ability to access computing resources on demand and pay-as-you-go with flexible and cost-efficient solutions. Fortunately, cloud environments are highly dynamic, but allocating resources efficiently can be difficult. Workloads may vary over time, and resources must be deployed to meet different performance targets while keeping within the stated budgetary allocations. This implies the tradeoff optimization of cost performance using predictive modeling, automation, and learning from diverse optimization methods. Thus, several solutions are given, such as criteria rules, reinforcement learning, metaheuristic, mathematical programming, and game theory. The aim is to choose the best resource mix that is economically viable and satisfies the threshold performance. There are plenty of benefits for the correct allocation, including less over-provisioning, better performance of applications, more efficient infrastructure use, and evidence-driven planning. The progress made is substantial. However, there still are significant difficulties around benchmarking strategies, uncertainties, coordination, business objectives, designs, and implementations that are robust and scalable. Further study is crucial to retrieve the economic profit of the cloud's elasticity fully. This article reviews the most recent research on the optimal allocation of resources on cloud computing platforms and cost efficiency.

***Corresponding author**

Ankur Mahida, Subject Matter Expert (SME), Barclays, USA.

Received: May 05, 2022; **Accepted:** May 11, 2022; **Published:** May 19, 2022

Keywords: Cloud Computing, Resource Allocation, Optimization, Cost Efficiency

Introduction

Cloud computing is an emerging revolutionary approach for IT infrastructure in that it allows on-demand access to a shared pool of virtualized resources through the internet. Cloud computing services can be categorized into three fundamental service models which include: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). One reason why the cloud model comprises of an edge is elasticity, which allows it to increase or decrease resource provisioning dynamically in real-time. The flexibility by itself can be the very reason for issues in the efficiency of resource allocation. Determining the best resource allocation, with numerous constant changes in workload, poses a challenge. Organizations seek to derive the highest performance within budget constraints. This scenario includes both under-provisioning and over-provisioning that affect application performance and incur costs that could have been avoided. Striking a balance of such a delicate nature necessitates the thoughtful optimization of resource allocations in line with the demands. This ensures that the needs are satisfied precisely and in the most cost-effective way. Developing strong optimization models for resource allocation is necessary if the cloud computing ecosystem is to achieve the economic potential it promises. The flexibility and affordability in computing do not imply sacrificing performance while utilizing efficient allocation mechanisms to match workloads and pricing that adapt to variability.

Problem Statement

The workload in cloud systems fluctuates drastically, resulting in significant differences in usage. Resource utilization, such as computing, storage, networking capacity, etc., needed by the application fluctuates with user traffic and resource usage. Load increase is usually observed more during the day and decreases at night. In addition to that, a sudden spike in demand is not an exception. Therefore, this dynamic characteristic will cause static allocation of resources to lead to over-allocative operations during low utilization and under-allocative operations when demand sharply grows. Allocation mechanisms must incorporate predictive modeling and adaptive real-time logic to size resources so they will be proportional to changing workloads.

The cloud providers deliver a heterogeneous mix of instances, sizes, configurations, and capabilities. Other examples, like general purpose, burstable, and memory-optimized, can go as far as GPU/FPGA-powered. Multipoint and hyperthreaded options provide choices. The question of how to simultaneously allocate among this diversified portfolio having distinct performance, cost, and reliability profiles introduces substantial complications to the decision-making process. There is a need to design the algorithms efficiently to match each application's tasks to the particular kind of heterogeneous resources.

Resource contention and interference can happen because cloud resources function as multi-tenant; hence, when multiple applications get to share a single physical infrastructure. Storage and cache retrieval will become issues due to congestion, thrashing, and I/O bottlenecks when allocations arise. It can lead to undesirable performance variability across co-located subsystems and application performance degradation. Algorithms must have a model of workload profiling and contention effects to maximize

the sharing with minimum interference.

For efficiency, cloud service providers pool customers into shared capacity by overbooking and reserving cache space. However, this also requires dealing with the fact that the resources will be short, and the service degradation occurs at the peak aggregate demand occasions. Thus, the optimization methodology must account for predictability in allocations through under-booking while providing room for cost-saving through over-provisioning. Robustness arrangements are the requirement to regulate shocks.

Solutions

Threshold-Based Rules

The most straightforward Traditional approach is using threshold-based rules based on metrics like CPU utilization being either upper or lower than given threshold values [1]. For instance, new instances will be deployed when the CPU averages 80% or more to meet additional demand. Ease implementation of the rules makes threshold-based rules a preferred choice. Yet, this is a list of the regulations that are static and reactive. They are weak on unpredictability and can be too conservative or aggressive as workload changes.

Reinforcement Learning

Reinforcement is a learning technique that has emerged as a promising tool for the online optimization of cloud resource allocation. A learning-based intelligent agent can optimize a policy by repeatedly going through a trial-and-error process with the cloud [2]. The agent performs the resource allocation actions and retrospectively checks the impacts on the cost and performance measures to adjust its allocation strategy. One of the vital strengths of the system is the ability to re-appropriate the allocations as optimal workloads and conditions as they vary over time.

Metaheuristic Algorithms

Since the cloud allocation problem is complex, we propose using metaheuristic algorithms, including genetic algorithms, simulated annealing, and ant colony optimization, to find near-optimal solutions to the problem [3]. These algorithms generate solutions through the exploratory iterative search for possible solutions using processes inspired by evolution, thermodynamics, and ant colonies. They can perfectly deal with large environments that have different resource types and applications. The catch is the increased computational intensity.

Mathematical Programming

Resource allocation optimization problem posed as a mathematical optimization problem permits deriving optimal solutions quickly. The problem is modeled using linear or nonlinear programming to reduce cost under constraints associated with performance [4]. Various solvers with high optimization power can be used to compute the most optimal allocation if workload forecasts are provided. Nevertheless, the complexity and scalability of problems are complex, especially for a public cloud environment.

Game Theory

Game theory models the bargaining between selfish cloud providers and users as non-cooperative or cooperative games. The strategy is to align decision-making on pricing and allocation between the two sides [5]. This methodology considers economic tradeoffs and competition on the cloud platform; however, it models human behavior and workloads under over-simplifying assumptions.

Uses

Efficiently using resources is a vital success factor for entities to get on board with the cloud and attain its cost-effectiveness. Optimization implies under-spending by narrowly matching allocated capacity to real-time load fluctuations, thus avoiding expensive oversubscription. Allocation does not allocate resources based on peak demand per se, but only those resources that are in use are provisioned when utilization varies [6]. This eliminates the cost of unutilized capacity during imbalanced periods of supply and demand. Research carried out by the industry suggests that optimizing infrastructure allocation can reduce cloud costs by 25% to 50%. This means millions of dollars saved from the computing power of hyperscale providers such as AWS and Azure [7].

In addition to cost reduction, effective allocation makes applications more performant and available by proactive resource provisioning so that the minimum required amount can be allocated as workloads change [8]. This avoids the issue of over-provisioning, where the utility needs to pay for the excess capacity necessary for spikes in demand. Mechanisms of optimization like predictive autoscaling analyze the user trends and predict the subsequent usage capacity. This diplomacy allows predictive allocations to be made to prevent resource shortages. Consequently, the users witness better service reliability with fewer drops in output quality.

In addition to this, the optimized allocations' data and insights are vital to our long-term planning and decision-making, which cover resourcing and procurement. A past collection of usage data across distinct workload and allocation modes will be used to predict future capacity and TCO on different instance types or cloud providers. It thus makes it possible to pick the best sizes, families, and numbers for one's cost/performance ratio needs.

It benefits the Infrastructure because of the efficiency and utilization through the multiplexing and workload consolidation of the various resources. This helps to maximize the resources, and more workloads can run on the capacity assigned at lower costs. Providers' profits increase by serving more customers, which they accomplish using fewer data centers.

Therefore, well-balanced resource allocation makes it possible to guarantee timely customer performance and stay within the predefined budgets. Pro-active resource allocation capability in a provider-controlled environment allows resource usage to be scaled up and down within resource constraints and IT requirements while keeping expenses low. This enables predictability of crucial key performance indicators according to contractual level.

Impacts

Resource allocation, however, impacts all levels, including cost reduction, performance increase, and strategic adjustments for cloud providers and clients. Effective allocation is beneficial regarding the cost decrease, and approximate figures are 25-50% based on real-life cases. Twitter estimated that its 50% cloud resource cost optimization brought the company's enormous cloud costs down by 50% of anticipated value, reaching millions of dollars in savings [9]. In addition, as reported by Google, successful management of its data centers effectively decreased production costs. The costs that are saved are by way of reduced over-provisioning, increased asset utilization, and consolidation of workloads. Effective distribution ensures that providers must refrain from buying and leaving idle additional capacity in their systems for those periods when demand is irregular. They use their existing set bandwidth more by statistical multiplexing and sharing

[10]. In addition, users ensure billing is minimized by using the needed instance types, sizes, and numbers to meet workload demands. Achieving strategic purchases can be carried out with optimized allocations that provide insights.

More than cost reduction, efficiency is needed for efficient resource allocation to perform well by allocating enough resources to deal with dynamic loads [11]. Google anticipated the increased consumption due to optimized resource allocation strategies like workload profiling, predictive scaling, and performance-based assignment. Users experience fewer or no outages or degradations. Detecting indicators early, allocation algorithms can automatically take action in anticipation of insufficient capacity. The fact that cloud-based services deliver elasticity by adjusting resources to workload improves the quality of service.

Furthermore, there is the advantage of optimizing usage that provides more room for applying and getting the most out of the cloud capabilities. Organizations can allow innovative initiatives to run optimally by using fine-grained abilities to allocate specialized resources like PGAs and memory-optimized instances [12]. Application utilization optimization allows for identifying trends in application use and spending, which is good for forecasting cloud migrations. Companies can then develop cloud-native architectures and software that help to use efficient resource utilization by correct allocation better. Other than that, cloud providers can also use allocation optimization facilities to differentiate their hosting platforms from each other.

Optimizing resource allocation gives rise to revolutionary changes that improve costs, performance, and cloud strategies for vendors and users. Excellent research remains to fully implement these benefits across public, private, and hybrid cloud platforms and applications. Although considerable strides have been achieved, the advantages can already be seen.

Scope

Benchmarking Diverse Techniques

Due to the abundance of diversity in the proposed fast optimization techniques, there is a clear need for systemic benchmarking and comparison to showcase its power across a range of workloads and metrics [13]. Test suites with beneficial workloads that have been standardized will help for the repeatable evaluation validation of techniques to determine their effectiveness. Measures such as costs, performance, scalability, and robustness are essential. These benchmarks will lead to the provision of guiding insights for amelioration.

Handling Uncertainty

Hence, optimization algorithms face considerable challenges from unpredictable parameters such as load fluctuation, user behavior dynamics, and spot price fluctuations in cloud environments. The recipes should be built using stochastic optimization, online learning, etc. techniques to optimize resource allocation given irregular conditions. Probability distributions allow modeling uncertainties. Ensemble forecasts or model predictive control are the other tools for handling variation.

Coordination Across Regions

Coordination mechanisms must be robust and spread over different geographical locations for systematizing allocation in disparate data centers [14]. Data locality, electricity cost, redundancy needs, and regulatory restrictions should be essential in deciding where to place the workload and how requests will be routed. Moreover, Hierarchical optimization with multilayer approaches looks like a promising guide. However, the problem of organizing allocation on a global scale has yet to be solved entirely.

Incorporating Business Goals

Shaping top-level business goals into optimization formulations, such as maximizing revenue, ensuring fairness, or meeting contractual obligations, is difficult. Recourse to goal-directed allocation using preference expression, linguistic techniques, or Pareto optimization shall be resorted to. It includes "allocation optimization with business objectives".

Scalability

The computational complexity of allocation optimization has the scalability to double the problem size, hence becoming intractable for cloud-scale environments. Scalability can be achieved by researching and developing approximation algorithms, decomposition approaches, and distributed designs. Consider the case of the hierarchical decomposition, where the problem gets partitioned.

Architecting for Optimized Allocation

Innovations in cloud platform architecture that can untie the knot of optimization allocation also have promise. Furthermore, incorporating flexibility, immediate metrics, and modularity can ameliorate allocation efficiency. The design of superior topologies for effective allocation starting from scratch is an unseen area.

Conclusion

As a requisite for maximizing the transformative potential of cloud computing about elasticity and cost efficiency, the efficient use of resources is a crucial factor. Since the invention of numerous sophisticated optimization techniques through research, we can develop reinforced learning methods such as machine learning approaches, mathematical programming formulation, and game-theoretic modeling. Notable advancement is evident, with the implementation of adoption phases showing tangible efficiency improvements, effective utilization, performance improvement, and cost reduction. Nevertheless, there are manifold questions on how to benchmark the diversity of methods, consider uncertainties, reach robustness, coordinate across data centers, and integrate business goals to make all that at the cloud scale. However, much work still needs to be done to cope with these intricate issues and to bring the potential of optimized cloud resource utilization to fruition. Achieving this vision calls for conscious innovation in prediction and adaptive optimization algorithms, goal-aware systems, scalable architecture, and cloud platforms ready for easy allocability. As research continues to refine the technology, the optimal utilization of the cloud can facilitate its capability to provide dependable computing services as they grow in demand and remain economical. The field remains full of opportunities, although some remain challenging and require significant attention. In the end, balanced use is the main thing for fully getting the potential of cloud computing.

References

1. Tang P, Li F, Zhou W, Hu W, Yang L (2014) Efficient Auto-Scaling Approach in the Telco Cloud Using Self-Learning Algorithm. Global Communications Conference https://www.researchgate.net/publication/300416953_Efficient_Auto-Scaling_Approach_in_the_Telco_Cloud_Using_Self-Learning_Algorithm.
2. Zorzi M, Zanella A, Testolin A, Grazia M, Zorzi M (2015) Cognition-Based Networks: A New Perspective on Network

Optimization Using Learning and Distributed Intelligence. IEEE Access 3: 1512-1530.

3. Kalra M, Singh S (2015) A review of metaheuristic scheduling techniques in cloud computing. Egyptian Informatics Journal 16: 275-295.
4. Qing H, Haopeng Z (2015) Optimal Balanced Coordinated Network Resource Allocation Using Swarm Optimization. IEEE Transactions on Systems, Man, and Cybernetics: Systems 45: 770-787.
5. Mohammad MH, Shamim Hossain M, Sarkar J, Huh EN (2012) Cooperative game based distributed resource allocation in horizontal dynamic cloud federationplatform. Information Systems Frontiers 16: 523-542.
6. Saraswathi AT, Kalaashri YRA, Padmavathi S (2015) Dynamic Resource Allocation Scheme in Cloud Computing. Procedia Computer Science 47: 30-36.
7. Yasrab R (2018) Platform-as-a-Service (PaaS): The Next Hype of Cloud Computing. ArXiv https://arxiv.org/abs/1804.10811v1.
8. Zhang H , Jiang G, Yoshihira K , Chen H (2014) Proactive Workload Management in Hybrid Cloud Computing. IEEE Transactions on Network and Service Management 11: 90-100.
9. Frost JJ, Sonfield A, Zolna MR, Finer LB (2014) Return on Investment: A Fuller Assessment of the Benefits and Cost Savings of the US Publicly Funded Family Planning Program. Milbank Quarterly 92: 696-749.
10. Townsend AM (2014) Smart cities: big data, civic hackers, and the quest for a new utopia. New York: W.W. Norton & Company https://ssir.org/books/excerpts/entry/smart_cities_big_data_civic_hackers_and_the_quest_for_a_new_utopia.
11. Turuk, Ashok K, Sahoo B, Addya, Sourav K (2016) Resource Management and Efficiency in Cloud Computing Environments. IGI Global https://www.igi-global.com/book/resource-management-efficiency-cloud-computing/163866.
12. Raj S (2015) Neo4j High Performance. Packt Publishing Ltd https://www.packtpub.com/en-in/product/neo4j-high-performance-9781783555154.
13. Bassini S, Danelutto M, Dazzi P (2018) Parallel Computing is Everywhere. IOS Press https://www.iospress.com/catalog/books/parallel-computing-is-everywhere.
14. Candela L, Castelli D, Coro G, Pagano P, Sinibaldi F (2013) Species distribution modeling in the cloud. Concurrency and Computation: Practice and Experience 28: 1056-1079.