User Context-Aware Resource Allocation

Wisdom Network as a Service

Winter 2022

Abstract

On-demand allocation of resources and the ability to pay for what you use is the primary advantage of cloud computing. However, this can be realized only when customers know the suitable configuration of resources required for an optimal user experience and cost-effectiveness. Overprovisioning of resources costs more. On the other hand, underprovisioning results in a degraded user experience. Therefore, to leverage the capability of the cloud, it's necessary to optimize the cost of resources while ensuring a reasonable QoS (Quality of Service). Resource provisioning across cloud providers keeping the ideal costs in mind is a complicated process and depends on a variety of parameters.

Furthermore, the umbrella of pricing plans such as on-demand, reserved, and spot (bidding) resources complicate the problem. The main problem with existing cost-optimization techniques is that they do not consider the end user's context. This graduate paper aims to optimize the cost of resources while guaranteeing an agreeable Quality of Service, achieved by taking the heterogeneous context parameters into account and popular pricing plans across cloud providers.

Keywords

Cloud Computing, Resource Allocation, Cost Optimization, Context-Aware Systems.

One Liner

Reduce Total Cost of Ownership (ToC) of IT resources while guaranteeing agreeable Quality of Service (QoS) by using context-aware resource allocation systems.

Introduction

The report of 2020 indicates that more than ten billion dollars have been wasted on unused resources due to improper resource management. Furthermore, an estimated nearly 40% of the cost of operation is because of over-provisions of resources. To emphasize the degree of over-provisioning, almost half of all the allocated resources had an approximate CPU usage of 4% or less.

With the lack of development in cost optimization techniques and increase in digitalization worldwide, small businesses are easily affected by over-provisioning of resources and incurring a high cost of operation.

The true colors of the problem are set to become a challenge and seen in the spotlight when the company scales to handle higher traffic. Hence, it is crucial to introduce a framework for better resource management for cost optimization.

Critical Analysis

At this point, we know that cost optimization is coupled to many heterogeneous parameters/factors, to mention a few, the type of application (example: CPU intensive, IO write-heavy or read-heavy, handling big data), budget constraints, and agreeable threshold of Quality of Service (QoS). To give an example, consider an application in the health sector, where the latency of the application has to be the minimum and hence takes a higher priority to maintain low latency. Another sector where latency is crucial is the multiplayer online gaming sector, where irregular user efficiency can change the outcome of the game. In such an application, the cost alone is not the primary objective.

Moreover, the scale of the applications is not static in most cases, and the process of improving resource management is a continuous process and refined over time. Hence, it's necessary to have a feedback loop to optimize the performance of the application and tweak the input parameters for a better outcome. Furthermore, the existing solutions and research papers concentrate on a particular cloud service provider or a specific type of pricing plan such as on-demand. Therefore, a key takeaway from this is to develop a resource management system that takes a combination of pricing plans for an optimal solution.

Business Model Canvas

So far, we know the theoretical view of the technique for cost optimization for resources, keeping QoS in mind and by considering pricing plans across cloud providers. So then, using the lean canvas template to refine the problem statement and propose a plausible, practical solution, we have:

Problem Definition

Over-provisioning resources and finding ways for cost optimization can be reduced to a scheduling and allocation problem. To develop a solution to improve compute resource usage parameters such as CPU, GPU, memory, and instance storage by considering various input user experience context-related and other crucial parameters in a system with a feedback loop to constantly improve the performance without compromising the Quality of Service and presuming that the resources required depends on the workload to run a set of tasks for a user, where each of these tasks can have different memory and CPU requirements.

Solution and Unique Value Proposition

The summary of the solution is as follows:

Introduce a middleware between the application and the resource management system, the functionality of the middleware is to track the user activity, such as the number of requests, time taken between the request and the response, frequency of requests, size of the payload, comparison of response payloads to determine the frequency of change in data; on the other hand, the inputs for the feedback loop such as the number of misses of the expected threshold for response time, number of complete and partial failures in a workflow and other unexpected behavior. During the development of the solution, it's important to maintain flexibility to add and remove these parameters, which are dependent on the application and various for other use cases.

Solution Deeper Dive

The unique proposition of the solution is the ability to work across pricing plans and cloud providers and hence developed considering the following:

1. Pricing Plans: The commonly used pricing plans across cloud providers are on-demand, reserved, and spot instances (bidding). Since different applications have different workloads, we choose a combination of pricing models for a user based on the gathered user context.
2. Support for different VMs: Cloud service providers offer different VM sizes with different CPU, memory, and IOPS (input/output operations per second). Hence the type of VM allocated depends on the workload of the user.
3. Response time thresholds: The threshold for agreeable latency for a task can have different priorities depending on the application and, at times, on the user (Premium users). Setting this to an optimal value takes a series of analyses of gathered workload information.
4. Resolving inconsistency: The model we develop typically has a budget constraint. Hence, the availability of resources to reach the required QoS may not go hand in hand. Therefore a model is developed considering multiple possibilities and how the problem is solved. One such example is degrading the performance for all or a cohort of users by a certain percentage when the budget limit has reached.

5. Budget Planning: As a company grows, so does the number of users. Hence, a predictive model is in place to estimate cost per user based on certain patterns and translate that to an optimal target budget over time (weekly or monthly).
6. Allocation and scheduling: Allocating the right VM type and price does not always result in an optimal cost. Considering a combination of pricing plans and allocating resources at the right time is essential to reduce the total operational cost. In fact, the allocation strategy can be as specific as allocating spot instances for the time being until a larger on-demand resource is available for processing.

To conclude the solution, optimizing the TOC (Total Operation Cost) while guaranteeing QoS (Quality of Service), the uncertainty of resource allocation and heterogeneous parameters are taken into consideration, and a model is developed to predict and allocate appropriate resources across pricing plans, including spot instances.

Literature Survey

The end goal is to reduce the TOC and maintain QoS while renting cloud resources from cloud providers. An OCRP (Optimal Cloud Resource Provisioning algorithm) was proposed based on demand and reserved allocation of resources and considering the workload and price fluctuations. However, the problem with the approach was specifying the number of on-demand VMs and failing to consider the use of a range of available VM types, not to mention the use of spot instances (resources allocated by bidding) was completely ignored [5].

Another research paper used a 2-phase approach using a swarm optimization and genetic algorithm. Still, again, the use of different types of VMs, the uncertainty of the cost and other parameters, and the use of different pricing plans was utterly neglected [4].

Another technique called the Load Level based Optimization for Virtual machine Allocation (LLOOVIA), the primary concern of this optimization technique was the tight coupling with the application type, making it challenging to generalize [6].

Existing Alternatives

Some of the built-in tools offered by popular cloud providers include:

1. Amazon Web Services Cost Management Console.
2. Microsoft Azure Cost Management.
3. Google Cloud Platform Cost Management Tool.

And numerous other third-party services that offer cost management software. However, in most cases, resource management is based on the usage of resources alone and hence fails to handle volatility and does not take the user experience into factor.

Downside

Although the use of context-aware middleware can significantly reduce the TOC (Total Cost of Operation), the downside is the overhead for resource allocation and concerns over the privacy of the data shared with the middleware. The privacy of an application is partially shared between the cloud providers and cloud consumers. Managing resources only on the hardware metrics of the resource has better flexibility over context-aware systems. Introducing middleware would increase the ownership and need for high security of data by the cloud consumer than the cloud provider. Furthermore, cloud providers do not have an incentive to offer the service, nor does it have the business context.

Unfair Advantage

Since the use of context-aware middleware for infrastructure management directly correlating to the user experience is mainly discussed and experimented in research papers and not fully put to use in practice. However, using a wisdom network as a service in practice for improving the results and directly relating to the reduced cost compared to the traditional cost management systems, the unfair advantage is being the first alternative in the industry. However, for an average consumer, it's still a cost management software, and hence, emphasizing the differentiating factor is important despite some of the downsides.

Target Customers

The potential customers, a specific target for wisdom network as a service, include small businesses with increasing growth rates, where the operation cost of the company is drastically growing, and resources make up most of the expenditure, and enterprise users offering software services already in the space of handling large scale, where the operational cost of resources directly affect the revenue model due to the reduced cost per customer. Finally, companies where the parameters such as latency and high availability of the service are critical, making resource management for optimal costs challenging.

Cost Structure and Revenue Streams

When it comes to pricing, the upper limit of the product is the average percentage of cost reduced after using the software. Hence, the pricing depends on the scale at which the customer is operating and the value addition on the existing cost optimization tools. Therefore, the pricing must be custom for enterprise users and a generic plan for small businesses, such as a monthly or yearly subscription. Lastly, for applications where data sensitivity is of high priority, such as financial applications, an option for a self-hosted variant with a one-time lump sum cost structure followed by a nominal fee for software maintenance and updates.

Conclusion

The main focus of the wisdom network as a service is the problem with the provisioning of resources from the user experience perspective. Cloud providers have different pricing plans, such as on-demand, reservation, and spot pricing. Considering a combination of these plans gives the flexibility of incorporating the advantages of each of these plans. Furthermore, considering different VM (Virtual Machine) types offered by cloud providers gives an advantage of better resource utilization. The problem with cost optimization is the constant handover between over-provisioning and under-provisioning and other uncertainties such as the workload, change in prices of on-demand resources, and volatile bidding charges for spot instances. Developing a reliable model to solve all of these uncertainties and using cloud services to their utmost potential gives us an advantage over other traditional cost optimization tools that primarily rely on the use of resources as a black box.

Next Steps

To develop a model to solve a real cloud resource provisioning problem across use cases. Furthermore, making the concept of user experience not specific to a user but also a cohort of users, it's often common to create a cohort of users within the company, such as the super users. They are considered regular/active users and are assigned a higher priority for customer retention. Therefore, a deeper analysis of the model is necessary for different types of applications. Similarly, improving the predictive model to find the correlation between the demand for workload and accuracy of resource allocation. Lastly, solving for long-term cost optimization is missed in the current models, where the highest cost saving is from deciding the reserved resources for a long term of at least one year. In other words, although the upfront cost increases, the long-term cost over months or even years is reduced.

Resources

[1] A. Abdelhadi and T. C. Clancy, "Optimal context-aware resource allocation in cellular networks," IEEE Xplore, Feb. 01, 2016. https://ieeexplore.ieee.org/document/7440640 (accessed Jan. 28, 2022).

[2] Salman, Haitham & Ali, Raniah & Thabit, Kawther. (2018). Study and Implementation of Resource Allocation Algorithms in Cloud Computing. 7. 591-594.

[3] C. Sieber, S. Schwarzmann, A. Blenk, T. Zinner, and W. Kellerer, "Scalable Application- and User-aware Resource Allocation in Enterprise Networks Using End-Host Pacing," ACM Transactions on Modeling and Performance Evaluation of Computing Systems, vol. 5, no. 3, pp. 1–41, Nov. 2020, DOI: 10.1145/3381996.

[4] AmolCAdamuthe, VK Bhise and GTT hampi. Solving resource provisioning in Cloud using GAs and PSO. In Engineering (NUiCONE), 2013 Nirma University International Conference on, pages 1–5. IEEE, 2013.

[5] Sivadon Chaisiri, Bu-Sung Lee, and Dusit Niyato. Optimization of resource provisioning cost in Cloud computing. IEEE Transactions on Services Computing, 5(2):164–177, 2012.

[6] QianZhu and Gagan Agrawal. Resource provisioning with budget constraints for adaptive applications in Cloud environments. In Proceedings of the 19th ACM International Symposium on High-Performance Distributed Computing, pages 304–307. ACM, 2010.

Assignment submitted by Adesh Nalpet Adimurthy, B00886154, adesh.nalpet@dal.ca