CSCI 5193 Assignment: Design Challenge

Cost Optimization: Context-Aware Resource Provisioning

Winter 2022

Overview

On-demand resource (Compute and Storage) allocation is one of the main benefits of using cloud (Infrastructure as a Service). For effective utilization of resources, the benefit of the cloud amplifies when the customer context-aware systems can decide the optimal size of the resources required and cost-effectively allocate resources. Thus, the problem reduces to an allocation and scheduling problem for cost optimization.

Why does the challenge exist?

As per the report from Gartner, in the year 2020 [1] [2], nearly $11 billion was wasted on idle cloud resources. Furthermore, based on the data from the users of ParkMyCloud, about 40% of the resources are over-provisioned, with an average CPU utilization of ~4% for computing devices, accounting for wastage of $6.6 billion, summing to a whopping $17.6 billion.

Especially for small businesses, spending time and human resources to refine the cloud resource utilization may not be the topmost priority. Furthermore, the problem is viewed in the spotlight when the businesses grow and scale the existing architecture. So without a doubt, solving this issue and making it a norm for better utilization of resources is crucial and ideal to work backwards, rather than the traditional approach.

Objective and Constraints

The objective is to find ways to reduce the Total Cost of Ownership (TCO)/Operational cost of businesses while meeting the necessary Quality of Service standards (QoS) by reducing the instances of over-provisioning resources. And prevent under-provisioning of resources resulting in a degraded user experience due to the lack of user context for resource management. In a sentence, to reduce the cost of operation and enhance certain aspects of user experience.

The major constraints to consider are:

Security Risk: One way to achieve Context-Aware Resource Provisioning is by introducing a middleware responsible for resource allocation based on the request/user context. Cloud providers treat security risks as a shared responsibility [3]. In this model, introducing a middleware for resource management would need the client to cover the security of what they put in it.

Customer Privacy: Privacy and Context-aware systems often don't go hand-in-hand; measures have to be taken to ensure that the context cannot be traced back to the original user and need explicit privacy policy agreements from users of the application.

Encryption: Another vital task is to consider the use of sensitive information in context-aware systems. Very similar to running search queries on encrypted logs that may contain sensitive information.

Resources

- Research and survey on performance improvements using context-aware systems [7].
- Resource management to prioritize time, QoS (Quality of Service), and cost [5] [6].
- Microservice architecture and protocols for low latency applications (Important for the middleware to be a HA service - Highly Available) [4].
- Loadtest frameworks to test the performance overhead of context-aware resource management systems with different levels of heterogeneity in context [7].
- Flexibility of choosing suitable resource pricing such as on-demand, reserved, scheduled, and bidding (Example: Pricing model in AWS).

Existing solutions

Cloud resource provisioning/management exists in different forms, (1) Cost-optimization provisioning aims to minimize operational and network costs and maximize revenue/profits. Here, the cost is counted as the primary objective and often may introduce a trade-off between cost and Quality of Service. (2) Time minimization provisioning, certain applications have strict expectations to complete within a specified time. The primary goal is to reduce the delay or execution time, (3) Energy consumption minimization provisioning, (4) Quality of service maximization provisioning, to name a few [4] [5] [6]. In other words, the solutions exist individually, tackling a specific problem.

Obstacles in Conceptualization

- Lack of standardization (pricing plans and multiple VM types) across cloud providers to bring in standardization.
- Developing heuristics for solving the overall model for decision making in context-aware systems.
- Developing a regression function to predict computation time and cost (finding the relationship between context patterns and resource metrics).
- Due to context heterogeneity, it is complex to build one model that fits all use-cases.
- Defining clear correlation between scheduling and allocation, without which, cost-optimization would be challenging to achieve.

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

References

[1] "2020 State of the Cloud Survey from Flexera," info.flexera.com. https://info.flexera.com/SLO-CM-REPORT-State-of-the-Cloud-2020 (accessed Jan. 30, 2022).

[2] B. Supernor, "10,000 Years of Data Says Your Server Sizing is Wrong," ParkMyCloud, Dec. 04, 2020. https://www.parkmycloud.com/blog/server-sizing/ (accessed Jan. 30, 2022).

[3] Mohamed Almorsy, John Grundy, and Ingo Muller. "An analysis of the cloud computing security problem." arXiv preprint arXiv:1609.01107, 2016.

[4] Qi Zhang, Quanyan Zhu, and Raouf Boutaba. Dynamic resource allocation for spot markets in Cloud computing environments. In Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on, pages 178–185. IEEE, 2011

[5] Artur Andrzejak, Derrick Kondo, and Sangho Yi. Decision model for Cloud computing under SLA constraints. In Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS), 2010 IEEE International Symposium on, pages 257–266. IEEE, 2010.

[6] Ivan Rodero, Juan Jaramillo, Andres Quiroz, Manish Parashar, Francesc Guim, and Stephen Poole. Energy-efficient application-aware online provisioning for virtualized Clouds and data centers. In International Conference on Green Computing, pages 31–45. IEEE, 2010.

[7] M. Tajvidi, "Cloud Resource Provisioning for End-users: Scheduling and Allocation of Virtual Machines," 2019. http://unsworks.unsw.edu.au/fapi/datastream/unsworks:70431/SOURCE02 (accessed Feb. 01, 2022).