Optimizing Resource Allocation in Cloud Computing Services Using Branch and Bound Algorithms

Angela Livia Arumsari - 13521094
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail (gmail): 13521094@std.stei.itb.ac.id

***Abstract*— Cloud computing has gained significant importance in the enterprise sector, offering faster and more profitable services. The adoption of cloud computing has been steadily increasing, with a majority of companies already using computing infrastructure or cloud-based applications. Multi-cloud platforms have become popular due to their numerous benefits, but they also pose challenges such as increased costs and management complexities. Optimal resource allocation is crucial for maximizing the benefits and addressing challenges in multi-cloud environments. This paper focuses on analyzing resource allocation strategies that prioritize cost minimization, utilizing the Branch and Bound algorithm. By optimizing resource allocation using this algorithm, companies can minimize their expenditure while achieving efficient resource utilization.**

Keywords—Cloud computing, Multi-Cloud, Resource Allocation, Branch and Bound

I. INTRODUCTION

Cloud computing has become a crucial part of enterprises. In recent years, companies have used the cloud computing paradigm to run various computing and storage workloads. The cloud offers faster and more profitable services. Based on a survey by IDG in 2020, cloud adoption levels that have held steady have accelerated in the last two years. In 2020, 81% of survey respondents reported already using computing infrastructure or having applications in the cloud, compared to 73% in 2018. Another 12% plan to adopt cloud-based applications in the next 12 months, and 6% plan to do so in the next 1 to 3 years.

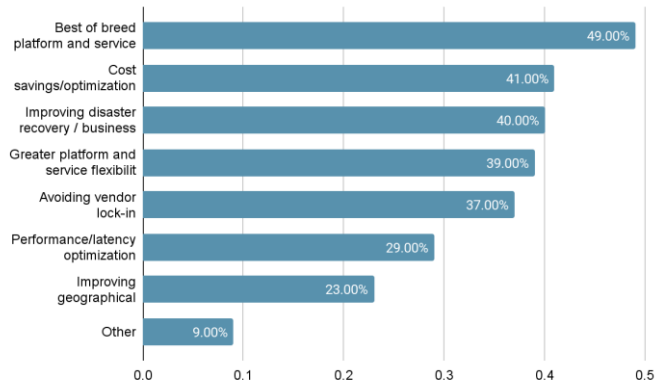

Fig. 1. Motivation in adopting multi-cloud platform. *(Source: [1] with modifications)*

More than half of respondents use multiple public clouds. Companies use multi-cloud platforms for several reasons as shown in Figure 1. Despite the many benefit of using a multi-cloud platform, the majority of respondents also report significant downsides to using a multi-cloud model [1]. One of the downsides is the increased costs due to cloud management and security challenges.

To tackle problems and maximize benefits in the multi-cloud platform, the services offered in terms of resources should be allocated optimally to the applications running in the cloud. Resource allocation in cloud computing involves scheduling and resource provision while keeping in view the available infrastructure, service level agreements, cost, and energy factors. Similarly, the resources have to be assigned in a way that every application gets the required resources without exceeding the limit of the cloud environment. In the same way, resource allocation is responsible for handling the issue of starving applications by properly resource allocation by enabling the service providers to allocate the resources for each module.

The issue of resource allocation is a significant challenge for cloud providers. The excessive consumption of resources has raised the need for better management. In addition, the resources required may exceed those available in the cloud as demand and capacity vary over time. Therefore, dynamic resource allocation techniques allow using the available capacity more efficiently.

In this paper, the focus is on analyzing resource allocation strategies that prioritize cost minimization using the Branch and Bound algorithm. The Branch and Bound algorithm is a powerful optimization technique that can be leveraged to optimize resource allocation in cloud computing. By employing this algorithm, companies can allocate resources in a manner that minimizes costs and maximizes efficiency. By minimizing costs, companies can effectively manage their budgets and allocate resources judiciously, leading to improved financial performance and operational efficiency.

II. BASIC THEORY

A. Cloud Computing

Cloud computing is on-demand access, via the internet, to computing resources that are hosted at a remote data center managed by a cloud services provider (CSP). The CSP makes these resources available for a monthly subscription fee or bills them according to usage.

Cloud computing has some advantages as follows.

- Lower IT costs

  Cloud offloads some or most of the costs and effort of purchasing, installing, configuring, and managing on-premises infrastructure.

- Improve agility and time-to-value

  Organizations can start using enterprise applications in minutes, instead of waiting weeks or months for IT to respond to a request, purchase and configure supporting hardware, and install software.

- Scale more easily and cost-effectively

  Cloud provides elasticity excess capacity that sits unused during slow periods, that can be scaled capacity up and down in response to spikes and dips in traffic.

Cloud computing has several types as follows.

- Public Cloud

  Public cloud is a type of cloud computing in which a cloud service provider makes computing resources available to users over the public internet. The public cloud provider owns, manages, and assumes all responsibility for the data centers, hardware, and infrastructure on which its customers' workloads run. It typically provides high-bandwidth network connectivity to ensure high performance and rapid access to applications and data. Some examples of the public cloud are Amazon Web Services (AWS), Google Cloud, IBM Cloud, Microsoft Azure, and Oracle Cloud.

- Private cloud

  A private cloud is a cloud environment in which all cloud infrastructure and computing resources are accessible by one customer only. A private cloud is typically hosted on-premises in the customer's data center. But a private cloud can also be hosted on an independent cloud provider's infrastructure or built on rented infrastructure housed in an offsite data center.

- Hybrid cloud

  A hybrid cloud is a combination of public and private cloud environments. A hybrid cloud connects an organization's private cloud services and public clouds into a single, flexible infrastructure for running the organization's applications and workloads. The goal of a hybrid cloud is to establish a mix of public and private cloud resources that gives an organization the flexibility to choose the optimal cloud for each application or workload and to move workloads freely between the two clouds as circumstances change.

- Multi-cloud and hybrid multi-cloud

  Multi-cloud is the use of two or more clouds from two or more different cloud providers. Organizations choose multi-cloud to avoid vendor lock-in, to have more services to choose from, and to access more innovation.

B. Resource Allocation in Cloud Computing

Resource Allocation (RA) plays a crucial role in cloud computing, involving the assignment of available resources to meet the demands of cloud applications. Precise management of resource allocation is vital to prevent service starvation. Resource provisioning addresses this challenge by enabling service providers to effectively manage resources for each module. The Resource Allocation Strategy (RAS) aims to integrate activities within the cloud environment to optimize resource utilization and allocation, aligning them with the needs of cloud applications. To achieve this, RAS requires detailed information on the type and quantity of resources required by each application to fulfill user jobs. The order and timing of resource allocation also contribute to an optimal RAS. An efficient RAS should address various criteria as follows.

- Resource contention situation arises when two applications try to access the same resource at the same time.
- Scarcity of resources arises when there are limited resources.
- Resource fragmentation situation arises when the resources are isolated
- Over-provisioning of resources arises when the application gets more surplus resources than the demanded one.
- Under-provisioning of resources occurs when the application is assigned fewer numbers of resources than the demand.

Both cloud users and providers have distinct roles in optimizing resource allocation. Cloud users provide estimates of resource demands and specify Service Level Agreements (SLAs) to ensure their job requirements are met. On the other hand, cloud providers supply information about resource offerings, current resource status, and available resources, enabling effective management and allocation of resources to host applications. The success of an optimal RAS is measured by parameters such as throughput, latency, and response time.

However, despite the reliability of cloud resources, dynamically allocating and managing resources across applications presents challenges. Cloud providers face difficulties in predicting the dynamic nature of user and application demands, making it impractical to anticipate resource requirements accurately. From the cloud user's perspective, completing jobs on time and minimizing costs are primary concerns. Therefore, due to factors such as limited resources, resource heterogeneity, locality restrictions,

environmental considerations, and the dynamic nature of resource demands, an efficient resource allocation system that caters to the unique characteristics of cloud environments is essential.

Resource allocation techniques employ various methodologies to effectively utilize resources and meet consumer requirements. In the context of cloud computing, these techniques can be categorized into six main categories: strategic, target resources, auction, optimization, scheduling, and power allocation. Each category encompasses different subheadings that further refine the techniques. The evaluation of these techniques is based on several parameters that are crucial from the perspectives of both cloud service providers and consumers.

For cloud service providers, cost is a significant parameter as it determines the expenses associated with providing different services. Resource utilization is another important consideration as providers aim to optimize the usage of resources to minimize idle capacity and overall data center expenditure. Power allocation is also critical in resource allocation strategies, given the increasing energy crisis. Minimizing power consumption aligns with environmental sustainability goals and helps make cloud services more eco-friendly.

From the standpoint of cloud service consumers, execution time and response time are crucial parameters. Both providers and consumers strive for minimal execution time for tasks, but it's important to note that executing multiple workloads on a single resource can lead to performance issues due to interference among the workloads. Therefore, response time, which measures the system's speed in responding to requests, becomes an essential metric for evaluating system performance.

User satisfaction is an important dimension, as cloud service providers aim to maximize customer satisfaction through effective resource allocation. By efficiently allocating resources in cloud computing, providers can enhance revenue generation and overall user satisfaction. Other parameters such as Quality of Service (QoS), Service Level Agreements (SLAs), fraud prevention measures, and revenue also play a role in evaluating resource allocation techniques.

These parameters are assigned values from 1 to 5, with 5 representing the highest value and 1 the lowest. However, it's important to note that the ideal value for each parameter may vary depending on the context. For example, in the case of cost, response time, execution time, workload, and power, lower values are considered ideal. Conversely, for user satisfaction, SLA, resource utilization, fraud prevention, and revenue, higher values are desirable.

Throughout the discussion, consistency has been maintained by following a consistent structure for each feature. This includes providing a definition of the feature, discussing relevant articles, and evaluating the feature based on the parameters mentioned above.

C. Branch and Bound Algorithm

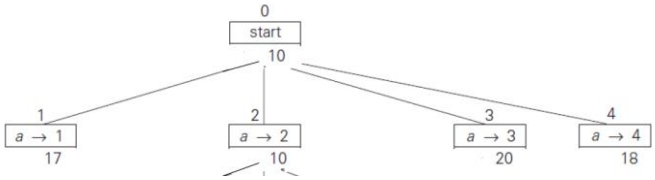

Fig. 2. Branch and Bound Algorithm. *(Source: [4] with modifications)*

Branch and bound is an algorithm design paradigm commonly utilized for solving combinatorial optimization problems with constraints. It follows a systematic approach of exploring the solution space by dividing it into smaller subspaces, known as branches, and efficiently pruning the search tree to avoid unnecessary exploration.

In the branch and bound algorithm, each node in the search tree is assigned a cost value, denoted as c(i), which represents the estimated cost of the cheapest path from the current node to the goal state. The algorithm aims to find the optimal solution by iteratively exploring the nodes with the smallest cost values, typically in the case of minimization problems.

To search for a solution, the branch and bound algorithm constructs a state space tree using the "best-first rule." This means that the next node to be expanded is not determined based on the order of its generation but rather on the node with the smallest cost value, indicating a potentially more promising path towards the goal.

The algorithm employs a bounding technique to efficiently eliminate unpromising nodes from further exploration. For each node in the state space tree, a method is employed to calculate the bound or best value of the objective function for each potential solution that can be obtained by adding components to the temporary solution represented by the node. By comparing the bounds of the nodes, the algorithm can prune branches that are guaranteed to yield suboptimal solutions, reducing the search space and improving computational efficiency.

This pruning process allows the branch and bound algorithm to discard entire subtrees that are not promising, thereby significantly reducing the number of solutions that need to be evaluated. As a result, it can efficiently handle problems with large solution spaces and constraints, enabling the exploration of only the most relevant portions of the search tree.

III. DISCUSSION

A. Limitations

There are many factors when applying the Branch and Bound algorithm to resource allocation in multi-cloud computing services. However, in this paper, the author set some limitations to make the calculation simpler. The limitations are listed below.

- Defining the cloud computing services without connecting with the cloud real services API.
- The restrictions taken into consideration are CPU, memory, and storage limit.
- The cost defined for every cloud service is the cost per virtual machine.

B. Cloud Computing Services and Virtual Machines Data

As mentioned in the limitations, the restrictions taken into consideration are CPU, memory, and storage limit. Therefore, for cloud computing services, the data that will be defined are CPU limit, memory limit, storage limit and cost per virtual machine. The data for computing services that will be defined is as follows.

TABLE I. CLOUD COMPUTING SERVICES DATA

Name	CPU Limit	Memory Limit	Storage Limit	Cost per VM
Service1	18	16	180	150
Service2	8	6	80	125
Service3	16	14	160	200

For the virtual machines, the data defined will be CPU capacity, memory capacity, and storage capacity. The data for virtual machines that will be defined is as follows.

TABLE II. VIRTUAL MACHINES DATA

Name	CPU Capacity	Memory Capacity	Storage Capacity
VM1	4	8	60
VM2	5	4	40
VM3	8	5	70
VM4	5	7	80
VM5	3	2	40

C. Branch and Bound Algorithm Implementation

To design a branch and bound algorithm for cloud computing resource allocation, some constraint needs to be defined. The constraint for the algorithm is as follows.

- CPU constraint: The total CPU capacity allocated to a service should not exceed the CPU limit of the service.
- Memory constraint: The total memory capacity allocated to a service should not exceed the memory limit of the service.
- Storage constraint: The total storage capacity allocated to a service should not exceed the storage limit of the service.
- Capacity constraint: Each virtual machine can be allocated to at most one service.

The objective function of the algorithm is to minimize the cost of cloud computing services. Therefore, for every node, the current lower bound for every node will be counted as the minimum cost for every virtual machine allocation in the available services. Because one service could be allocated to more than one virtual machine, the minimum cost would be the number of virtual machines times the cost per virtual machine in the lowest-cost services. The logical reason behind this bound is for every legitimate solution the real cost would be greater or equal to the lower bound.

Any available cloud computing services are available until the CPU, memory, or storage is full. The search for the solution will be using the Best-First search using the lower bound that has been described above. To make the algorithm more effective, the services will be sorted based on the ascending order of the cost per virtual machine.

The complete steps for the branch and bound algorithms are as described below.

- Sort the services based on the cost per VM in ascending order.
- Initialize the priority queue as an empty list.
- Create an initial solution with empty allocations and a cost of 0. Push this initial solution into the priority queue.
- While the priority queue is not empty: Pop the solution with the lowest cost from the priority queue. If the cost of the popped solution is greater than or equal to the best solution's cost, continue to the next iteration.
- Check if the solution is complete. If the solution is complete and the best solution is already found, continue to the next iteration. If the best solution is not set or the popped solution has a lower cost than the best solution, update the best solution to the popped solution first.
- Select the next service to allocate based on the partial solution and the sorted services list.
- Generate child solutions by branching to the next service. For each available VM that can be allocated to the next service, check if the allocation is feasible using the constraint that has been defined. If feasible, create a child solution by updating the allocations and cost. Append the child solution to the child solutions list. Check feasibility for the child solution using the constraint function. If not feasible, discard the child solution.
- Push the child solutions into the priority queue.
- Return the best solution.

D. Cost Optimization Analysis

For the available data, the first iteration from the root using Branch and Bound algorithms will look like follow.

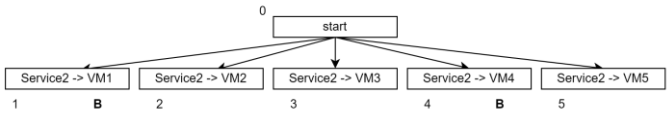

Fig. 3. First Iteration Branch and Bound Algorithm. *(Source: Primary)*

Node 1 and node 4 are being "killed" because the node violates the constraint that has been defined. For node 3, the allocation of Service2 made Service2 unavailable because the CPU is full. The lower bound cost is computed below.

TABLE III. NODES COST IN THE FIRST ITERATION

Node	Cost
2	625
3	725
5	625

If all the iterations from the node are violating the constraint, then the service can be assumed as unavailable and iteration continues to the next service. For resource allocation, the optimized solution could be found from different nodes because one service could be allocated to several virtual machines. Therefore, when there is a goal node with the other node having a cost not lower than the goal node, the search should stop.

To find the most optimal solution, the branch and bound algorithm should be continued until the solution is complete or all virtual machines have been allocated to the services. To know whether the solution is optimal or not, we need to see the cost of every node. If all of the nodes beside the current node has a cost higher than the current node, than the current node is the most optimal solution.

If the branch and bound algorithms are continued to the data above, the results will be as follows.

TABLE IV. VIRTUAL MACHINES DATA

Service	Virtual Machine Allocated	Cost per Service
Service1	VM1, VM3	300
Service2	VM2, VM5	250
Service3	VM4	200
Total Cost		750

By using a branch and bound algorithm, the most optimal solution is guaranteed to be found. In this case, the minimum cost for cloud computing services resource allocation is found.

IV. CONCLUSION

In conclusion, this paper highlights the potential of the Branch and Bound algorithm in optimizing resource allocation in cloud computing. The algorithm proves effective in finding the resource allocation strategy that minimizes the cost of services, which is a crucial objective for companies operating in the cloud environment. By optimizing resource allocation, companies can enhance their management of multi-cloud platforms, ensuring efficient utilization of resources while keeping costs under control.

While this research focused on a simplified version of the multi-cloud management platform due to limitations in time and resources, there is scope for further development. One possible avenue for future research is to incorporate real-time cloud services APIs, allowing for a more comprehensive evaluation of resource allocation strategies. By using real machines and cloud services, the algorithm's effectiveness can be tested and validated in practical scenarios, catering to the specific needs of companies operating in the cloud.

VIDEO LINK AT YOUTUBE

The video explanation of the paper can be found below.

https://bit.ly/VideoStima13521094

ACKNOWLEDGMENT

This paper would not have been possible without the plenty of resources on Algorithm Strategy provided by all the lecturers of IF2120. The author would also like to especially thank Dr. Nur Ulfa Maulidevi, S.T., M.Sc. as the lecturer of class 02, for assigning this paper, as it is a way to learn the applications of the materials taught in class.

REFERENCES

[1] A. Razzaq, A. Z. Abbasi, and A. Shahzad, "Resource Allocation Techniques in Cloud Computing: A Review and Future Directions," Elektronika ir Elektrotechnika, vol. 26, no. 6, pp. 40-51, Dec. 2020. [Online]. Available: https://doi.org/10.5755/j01.eie.26.6.25865. [Accessed May. 21, 2023].

[2] IBM, "Cloud Computing," [Online]. Available: https://www.ibm.com/topics/cloud-computing. [Accessed May. 21, 2023].

[3] "2020 Cloud Computing Executive Summary," [Online]. Available: https://cdn2.hubspot.net/hubfs/1624046/2020%20Cloud%20Computing%20executive%20summary_v2.pdf. [Accessed May. 21, 2023].

[4] "A Survey on Resource Allocation Strategies in Cloud Computing," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 3, no. 6, 2012. [Online]. Available: www.ijacsa.thesai.org. [Accessed May. 21, 2023].

[5] R. Munir, "Branch and Bound Bagian 4," IF2120 Strategi Algoritma. [Online]. Available: R. Munir, "Graf Bagian 1," IF2120 Matematika Diskrit. [Online]. Available: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2020-2021/Graf-2020-Bagian1.pdf. [Accessed May. 21, 2023].

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 22 Mei 2023

Angela Livia Arumsari 13521094