

AILEC3 - Artificial Intelligence for Lower Energy Consumption in Cloud Computing

Ettore Tancredi Galante
Department of Computer Science
University of Milan
Milan, Italy
ettoretancredi.galante@studenti.unimi.it

I. ABSTRACT

II. INTRODUCTION

The increasing adoption of cloud computing services in the last years has led to a significant increase in the energy consumption of data centers. Cloud computing providers are facing the challenge of reducing energy consumption while maintaining or improving the performance of their services. The scale of cloud computing infrastructures, the dynamic nature of workloads and the need to provide 24/7 availability make energy efficiency a critical concern. The energy consumption of data centers is expected to grow exponentially in the coming years, driven by the increasing demand for cloud services. This is not only a concern from an environmental perspective, but it also has significant financial implications for the companies that operate and use these services. Therefore, it is crucial to develop energy-efficient solutions in the field of cloud computing in order to mitigate the environmental and financial impacts of the increasing energy consumption of data centers.

The optimization of resource allocation in cloud computing environments is a complex problem that involves a large number of variables and constraints. One of the main objectives is to minimize energy consumption while maintaining or improving the performance of the services. The use of Artificial Intelligence (AI) techniques, such as genetic algorithms, has been widely studied as a method for addressing this problem. Genetic algorithms are a type of optimization algorithm inspired by the principles of natural selection and genetics. They use heuristic techniques to simulate the process of natural evolution in order to find near-optimal solutions for complex problems. Genetic algorithms are particularly well suited for solving problems that involve a large number of variables and constraints, such as the resource allocation problem in cloud computing. They have the ability to explore a large search space in a relatively short amount of time, allowing them to find near-optimal solutions even in the presence of incomplete or uncertain information. This is important in cloud computing, where the workload and resource usage patterns are often dynamic and difficult to predict. The use of genetic algorithms in cloud computing has the potential to significantly improve the energy efficiency of data centers while maintaining or improving performance.

One of the key aspects of measuring energy consumption in cloud computing is identifying the power consumption of the various components of the system, such as servers, storage devices, and network devices.

These components consume energy in the form of electricity, which is measured in watt-hours (Wh) or joules (J). The energy consumption of a data center can be measured at different levels of granularity, from individual servers to entire racks or data centers. One of the most commonly used metrics for measuring energy consumption in cloud computing is Power Usage Effectiveness (PUE).

PUE is a ratio of the total energy consumed by a data center to the energy consumed by the IT equipment. A PUE of 1.0 indicates that all of the energy consumed by the data center is used by the IT equipment, while a PUE of greater than 1.0 indicates that some of the energy is being used for other purposes, such as cooling and lighting.

As the number of IPS executed by a process on a server increases, the energy consumption of the server also increases. This, in turn, can lead to an increase in the PUE of the data center. This is because the additional energy consumed by the server must be taken into account when calculating the PUE. Therefore, as the number of IPS increases, the PUE will also increase, indicating a decrease in the overall energy efficiency of the data center.

The optimization of resource allocation in cloud computing environments is a complex problem that involves a large number of variables and constraints. These variables include the number of servers, storage devices, and network devices, as well as the workloads that are running on these devices. The constraints include energy consumption, performance, and availability. Artificial Intelligence (AI) techniques, such as genetic algorithms, have been widely used to address this problem. Genetic algorithms are a type of optimization algorithm inspired by the principles of natural selection and genetics, which use heuristic techniques to simulate the process of natural evolution in order to find near-optimal solutions for complex problems.

Genetic algorithms are particularly well suited for solving problems that involve a large number of variables and constraints, such as the resource allocation problem in cloud computing. This is because they have the ability to explore a large search space in a relatively short amount of time.

The search space in the case of resource allocation in cloud computing is the set of all possible combinations of resources that can be allocated to the different workloads. The large size of this search space makes it impractical to use traditional optimization techniques, such as exhaustive search or gradient-based methods, which are unable to explore the entire search space in a reasonable amount of time.

Additionally, genetic algorithms are able to find near-optimal solutions even in the presence of incomplete or uncertain information. This is important in cloud computing, where the workload and resource usage patterns are often dynamic and difficult to predict. The ability of genetic algorithms to adapt to changes in the workload and resource usage patterns makes them well suited for dealing with the dynamic nature of cloud computing environments.

The use of genetic algorithms in cloud computing has the potential to significantly improve the energy efficiency of data centers while maintaining or improving performance. This is because genetic algorithms can be used to optimize the resource allocation in a way that minimizes the energy consumption while maintaining the performance and availability of the services. The optimization of resource allocation can lead to a reduction in the number of servers and storage devices that are required to run a given workload, which in turn, reduces the energy consumption. Additionally, genetic algorithms can be used to adapt to changes in the workload and resource usage patterns, which can further improve the energy efficiency of the data center.

Despite the advantages of using genetic algorithms for resource allocation in cloud computing, there are also some limitations to consider. One of the main limitations is the computational expense of genetic algorithms, particularly when the problem size is large. The resource allocation problem in cloud computing environments can involve a large number of variables and constraints, which can make the computational cost of genetic algorithms quite high. This can make them impractical for certain types of problems or when resources are limited. Additionally, genetic algorithms are often sensitive to the choice of parameters and initial conditions. This can make it difficult to obtain consistent and reliable results, as the choice of parameters and initial conditions can have a significant impact on the performance of the algorithm.

The selection of suitable parameters and initial conditions is an important step in the implementation of genetic algorithms. It requires a significant amount of expertise and experience to choose the appropriate parameters and initial conditions that will lead to good performance. Furthermore, the results obtained from genetic algorithms are probabilistic in nature, which means that multiple runs of the algorithm are typically required to obtain a stable and robust solution. This can increase the computational cost and time required to obtain a solution, which can be a limitation when resources are limited.

In addition to computational expense and sensitivity to parameters, genetic algorithms also have some other limitations such as the lack of guarantee of global optima, and the risk of getting stuck in local optima. Genetic algorithms are

also not suitable for problems with deterministic solutions. Furthermore, the stochastic nature of genetic algorithms may also lead to a lack of reproducibility of results.

In conclusion, it is crucial to carefully evaluate the trade-offs between the potential benefits and limitations of using genetic algorithms in cloud computing before deciding to implement them. While genetic algorithms can be an effective method for optimizing resource allocation in cloud computing environments, their computational expense, sensitivity to parameters and initial conditions, lack of guarantee of global optima and stochastic nature, among other limitations, must be taken into account when deciding to implement them. It is important to consider the specific characteristics of the problem at hand, and weigh them against the computational resources and expertise available, before deciding to use genetic algorithms in cloud computing environments.

III. STATE OF THE ART

IV. DATASET

V. BASELINE

VI. EXPERIMENTS

VII. RESULTS

VIII. CONCLUSIONS

IX. APPENDIX

REFERENCES