IJFEAT
INTERNATIONAL JOURNAL FOR ENGINEERING APPLICATIONS AND TECHNOLOGY

Optimal Resource Allocation approach in Cloud Environment for current Trends

ANIL,
Research Scholar, School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh, India
anil.2020@vitbhopal.ac.in

Dr. H. AZATH,
Senior Assistant Professor Grade 2, School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh, India
azath.h@vitbhopal.ac.in

Dr. Muneeswaran V,
Senior Assistant Professor, School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh, India
muneeswaran@vitbhopal.ac.in

Abstract: This article describes the job of the resource allocator, the factors that are considered when assigning resources, and the resource allocation processes. The utility-based distribution of resources to clients is made possible by cloud computing. The system for allocating resources must consider the optimal usage of resources and the lowest possible cost per customer service. This condition for resource allocation assumes that resource allocation will be considered an optimization issue. To conduct the optimal distribution, the distributor must be armed with information about the current state of the properties, the anticipated demand from consumers, and the dynamic variations in the cloud. This study has classified the resource allocator's parameters according to the allocation cost, resource usage, and time considerations.

Keywords: Deep Learning, Cloud, Resource Allocation Algorithms.

1. Introduction

The cloud is a collection of scattered and linked resources accessible over the Net. In cloud computing, capitals like CPU, memory, storage, and bandwidth are communal by assigning Virtual Machines (VM) to customers. According to the needs of the user application, the VMs are outfitted with the necessary possessions. The cloud delivers commodity-based utility resources, enabling the cost-effective delivery of these resources. This allocation on a per-demand basis delivers a cost-effective resource supply. The allocation of virtual machines is governed by the service level contract between the service benefactor and the customer. The cloud supply allocator manages the heterogeneous capitals and the demand for capitals that vary following many consumer apps. The supply provisioning in the cloud handles the scalability of the workload's active fluctuations. Cloud computing gives users adaptability depending on the availability of applications and resources. Providing an active method for an adaptive and effective

resource allocation instrument is a problem posed by the different resource kinds and uses [1-4].

The goal of adequately allocating and efficiently using resources provides the resource manager with a substantial problem. The environment is comprised of a range of users and applications with fluctuating resource needs. The dynamic approach to resource allocation is required to provide a speedier and more equitable distribution of resources and map natural resources to virtual resources given to a user [3]. Numerous articles on the subject of resource allocation have been written. Power management and performance efficiency are the most critical factors when allocating resources. Other considerations during resource allocations (RA) include the resource type, resource parameters, scheduling techniques, and cloud architecture. The resource allocator must be aware of the state of the resources to allocate them dynamically. Observations indicate that addressing the RA issue as an optimization problem improves resource utilization efficiency due to real-time demand changes and resource availability.

2. Related Work

The resource allocator has evolved into a crucial constituent of cloud supply management. The cloud application, facilities, and capitals are dispersed with diverse goals in the cloud atmosphere. The cloud atmosphere contains abundant capital such as processing components, memory, and storage for operator requests, the demands of which are scaled up or down based on demand. The prime objective of resource distribution techniques is to increase cloud resource providers' revenue by maximizing resource use. From the user's perspective, the goal is to satisfy expectations at the lowest possible cost.

Throughout the research, the policy, approach, and parameters for supply distribution in cloud computing are given to the policy. Separated from one other were the review articles and the implementation papers. In this part, the review articles are analyzed. In sections 3 and 4, the review of other papers is covered.

The writers of [5] review the mechanics of RA. A prototypical classification is presented. The writers have evaluated data that considers the investigation gap between known instruments and prospective study themes. The cataloging is based on instrument type as static or active, processing manner as centralized or dispersed, QoS as in resource consumption, energy consumption, response time and charge, purposes as in sole objective or numerous objectives, Service Level Arrangement (SLA)-based as in adaptive or non-adaptive, policy-based as in implementation time, VM-based, gossip, utility-based, auction or application-based, method of assessment as in execution or replication, and application type.

The authors of reference (6) present an overview of RA in the cloud and classify allocation approaches as agent-oriented, priority-based, quality-based, active, and QoS-based. Cloud elasticity is analyzed in

[7]'s work.

3.Resource Allocation Parameters

The QoS criteria are selected by the customer and the cloud service provider, and they are included in the SLA. While assigning resources, the RA will assess the restrictions and the accessibility of possessions and then assign the capitals. Several works on the mentioned parameters are found in the literature. The importance of the resource allocation policy cannot be overstated [8].

This part has analyzed the factors that may be evaluated and the numerous RA policies accessible in the literature. The supply needs may be characterized as network supplies, such as bandwidth, latency, and throughput, and computing supplies, such as CPU and remembrance [9]. Few writers have considered any of the factors for RA, while some have considered more than one element when determining the most appropriate supply to an application. They are given as both single- and multiple-objective RAs.

The adopted RA strategy must minimize the resource distribution cost, the total system operation, and the task implementation time. These reflections bring us to the RA as an optimization issue to decrease the total price while incorporating the concept of boosting the complete dependability. According to the dependability examination, failure is one of the RA's reliability metrics. The recommended RA techniques are based on the cost of resource distribution, system operation, task implementation time, and dependability.

4. Resource Allocation Policy

Regardless of the criteria used to display the supply request, the RA strategy or the stated approach would assure the best distribution of resources. This segment includes a literature overview on resource allocation policies and strategies. Dynamic supply-demand exists in the cloud and also requires dynamic allocation mechanisms.

In resource allocation and scheduling, machine learning eliminates human intervention. The system allocates resources optimally, which might increase the profit of cloud service providers. In order to do cataloging based on the SLA and presentation cost, linear regression and Bayesian networks are applied. The Support Vector Machine (SVM) is used to differentiate between apparatuses that have been detached due to letdown or preservation and fully functional ones. Classification takes into account CPU and memory use. The best allocation technique is chosen based on the fitness value to prevent distributing unusable or unreliable resources. Where the fitness value is calculated based on the average time between letdowns. In auction-based resource allocation, linear and logistic regression calculates the auction and determines the optimum allocation option.

The Long Short-Term Memory (LSTM) prototypical is used to anticipate the supply use of each program, taking into account CPU and remembrance usage. The link between demanded and used capitals is

analyzed, and the execution alleviates the overallocation and under-allocation of capitals. The Auto-Regressive Integrated Moving Average (ARIMA) model was also developed to forecast workload prediction and its influence on QoS. It is a practical strategy for predicting workloads dynamically. The autocorrelation between assignment request and assignment is used, and the assignment analyzer determines the new changes in workload.

The virtual machines are allocated using a general mechanism known as the threshold-based tournament selection method. For the implementation, a genetic algorithm is employed. The request scope is condensed to the maximum training size of the computer-generated machine or the supply and transformed into binary to create a gene. Also transformed into a gene is the processing power of the virtual computer. The chromosome is produced by combining a binary application with a randomly selected, binary-to-binary transformed virtual apparatus. The winner of the threshold-based contest is determined by a competition between two randomly chosen persons. As the most suitable resource for allocation, the resource matching the winner is chosen.

Learning based on reinforcement learning (RL) considers various aspects of all accessible data sources. This is a good quality for resource autonomy management. Each parameter is evaluated with the value function, and the function value and related action are to be placed in the lookup table. To build the learning environment, artificial neural networks are deployed. Neural networks can remember value functions without extra storage, which is a distinct benefit. The restriction of this effort is that the state of the capitals is determined, and the capitals are assigned to the requesting robots only if the request can be satisfied. Otherwise, local processing is performed by the robots.

The RL-based multi-object optimization technique improves RA energy use and reduces power consumption. Presentation and energy are compromised, and SLA breaches may occur throughout the procedure. Consideration is given to power usage and SLA breaches to enhance performance. Observing energy use enables more significant energy conservation. The algorithm selects the ideal pair of request and host.

The values of Q are recorded in a table in RL. It becomes a design difficulty for real-world requests to store all the incessant state space and action space on the table. In order to address this issue, function calculation is used to excerpt features from representations, value functions, or rules, and then deep neuronal networks (DNN) are employed to calculate the whole function. The RL evaluates evolutions based on the likelihood of accumulating greater rewards. Deep reinforcement learning (DRL) explores when learning chances are valuable enough for the mediator to pursue without a separate training phase. The bulk of DRL procedures is model-free and appropriate for situations that cannot be represented. Using these function approximators, one may extract the

features.

5. Conclusion

This paper presents research on options for resource distribution. Cloud resources' scalability and on-demand nature make resource allocation an optimization challenge. The paper discusses the criteria that are evaluated when allocating capital and the various methodologies for allocating resources. There is evidence of work on RL and DRL in the literature, the bulk of which considers power management and energy to be the core factors of interest.

References

[1] Souravlas, S. and Katsavounis, S., 2019. Scheduling fair resource allocation policies for cloud computing through flow control. Electronics, 8(11), p.1348.

[2] Souravlas, S., Katsavounis, S. and Anastasiadou, S., 2020. On Modeling and Simulation of Resource Allocation Policies in Cloud Computing Using Colored Petri Nets. Applied Sciences, 10(16), p.5644.

[3] Fard, M.V., Sahafi, A., Rahmani, A.M. and Mashhadi, P.S., 2020. Resource allocation mechanisms in cloud computing: a systematic literature review. IET Software, 14(6), pp.638-653

[4] Bharanidharan, G. and Jayalakshmi, S., 2021, March. Elastic Resource Allocation, Provisioning and Models Classification on Cloud Computing A Literature Review. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1909-1915). IEEE

[5] Duggan, M., Mason, K., Duggan, J., Howley, E. and Barrett, E., 2017, December. Predicting host CPU utilization in cloud computing using recurrent neural networks. In 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST) (pp. 67-72). IEEE

[6] Sathiyamoorthi, V., Keerthika, P., Suresh, P., Zhang, Z.J., Rao, A.P. and Logeswaran, K., 2021. Adaptive Fault Tolerant Resource Allocation Scheme for Cloud Computing Environments. Journal of Organizational and End User Computing (JOEUC), 33(5), pp.135-152

[7] Shabka, Z. and Zervas, G., 2021. Nara: Learning Network-Aware Resource Allocation Algorithms for Cloud Data Centres. arXiv preprint arXiv:2106.02412.

[8] Mehmood, T., Latif, S. and Malik, S., 2018, October. Prediction of cloud computing resource utilization. In 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT) (pp. 38-42). IEEE

[9] Liu, D., Sui, X., Li, L., Jiang, Z., Wang, H., Zhang, Z. and Zeng, Y., 2018. A cloud service adaptive framework based on reliable resource allocation. Future Generation Computer Systems, 89, pp.455-463.

[10] Zhang, J., Xie, N., Zhang, X., Yue, K., Li, W. and Kumar, D., 2018. Machine learning based resource

allocation of cloud computing in auction. Comput. Mater. Continua, 56(1), pp.123-135.

[11] Thonglek, K., Ichikawa, K., Takahashi, K., Iida, H. and Nakasan, C., 2019, September. Improving resource utilization in data centers using an lstm-based prediction model. In 2019 IEEE International Conference on Cluster Computing (CLUSTER) (pp. 1-8). IEEE.

[12] Saxena, D., Singh, A.K. and Buyya, R., 2021. OP-MLB: An online VM prediction based multi-objective load balancing framework for resource management at cloud datacenter. IEEE Transactions on Cloud Computing

[13] Tournaire, T., Castel-Taleb, H. and Hyon, E., 2021. Optimal control policies for resource allocation in the Cloud: comparison between Markov decision process and heuristic approaches. arXiv preprint arXiv:2104.14879

[14] Liu, H., Liu, S. and Zheng, K., 2018. A reinforcement learning-based resource allocation scheme for cloud robotics. IEEE Access, 6, pp.17215-17222.

[15] Thein, T., Myo, M.M., Parvin, S. and Gawanmeh, A., 2020. Reinforcement learning based methodology for energy-efficient resource allocation in cloud data centers. Journal of King Saud University-Computer and Information Sciences, 32(10), pp.1127-1139.

[16] Liu, N., Li, Z., Xu, J., Xu, Z., Lin, S., Qiu, Q., Tang, J. and Wang, Y., 2017, June. A

[17] hierarchical framework of cloud resource allocation and power management using deep reinforcement learning. In 2017 IEEE 37th international conference on distributed computing systems (ICDCS) (pp. 372-382). IEEE.

[18] Zhang, Z., Zhang, D. and Qiu, R.C., 2019. Deep reinforcement learning for power system applications: An overview. CSEE Journal of Power and Energy Systems, 6(1), pp.213-225.

[19] Shrimali, B. and Patel, H., 2020. Multi-objective optimization-oriented policy for performance and energy-efficient resource allocation in cloud environment. Journal of King Saud University-Computer and Information Sciences, 32(7), pp.860-869.

[20] Souravlas, S., Katsavounis, S. and Anastasiadou, S., 2020. On Modeling and Simulation of Resource Allocation Policies in Cloud Computing Using Colored Petri Nets. Applied Sciences, 10(16), p.5644.