A MULTI-STAGE RECURRENT NEURAL NETWORKS FRAMEWORK FOR IMPROVING RESOURCE ALLOCATION IN CLOUD

S. Vimalnath
Department of Computer Science and Engineering, Paavai Engineering College, India

Abstract

The study strives to expand the virtualization technique to create a method that employs efficient architecture and algorithm for enhancing the flexibility and the resource allocation effectiveness of future cloud computing. This method uses Recurrent Neural Networks (RNN) to improve the resource allocation. For this purpose, simulation is done and results are obtained using the GreenCloud Simulator software. The results show that the proposed framework is 20% more efficient than the current QoS improvement framework.

Keywords:
Cloud Computing, Quality of Service, Virtualization

1. INTRODUCTION

A cluster of servers or computers are connected to create a cloud to provide users with resources. It is described as an attractive resource pool which users can use through the Internet. The key feature of Cloud computing is distribution, virtualization, and dynamic expansion. The next step to the development of the Internet is cloud computing.

Cloud computing has many points of interest, with the most important being lower costs, re-supply of assets and remote access. Organizations can avoid the investment in physical IT framework by renting the services from third party cloud service providers. Due to the elasticity of cloud computing, we must rapidly use numerous resources from cloud providers when developing our business. The services will be delivered with the best resources for the application in the cloud [2].

1.1 ARCHITECTURE OF CLOUD

In essence, there are two components in cloud computing namely, the front end and the rear, as shown in Fig.1. The front end of Cloud computing is defined as the client. It includes interfaces and applications which the Cloud Computing Platform is expected to access. It is said that the Cloud itself is the back end of cloud computer systems.

Cloud Computing Services provide resources that include virtual machines (VM), servers, data storage, security systems, and much more. This distributes the file system across numerous hard disks and machines that ensure data recovery even in case of a system failure. If the data fail at one place, the other routinely takes over. The distributed file system is dispensed on the client disk space, while another key part is the algorithm for the assignment of resources. Intensive algorithms [3]-[4] are used in cloud computing as a strong scattered environment.

1.2 RESOURCE ALLOCATION

The cloud computing resources are the source of the required computing necessities. The resources are divided into two types: hardware and software resources. The Data Centers and servers are the hardware resources, while client- and server-side software, information and application-related resources are the software resources [5]. The allocation of resources as data center management, networking and operating systems is examined in many computing regions.

Any framework that expects to provide the applications needed by the framework of the supplier may also be considered as a Resource Allocation System (RAS) in Cloud Computing. In addition to confirming the developer, resource allocation mechanisms should also take into consideration the current status of all the resources in the cloud environment, taking into account the final goal of influencing algorithms to provide developers with both physical and virtual resources, thus reducing the operating cost of the cloud environment.

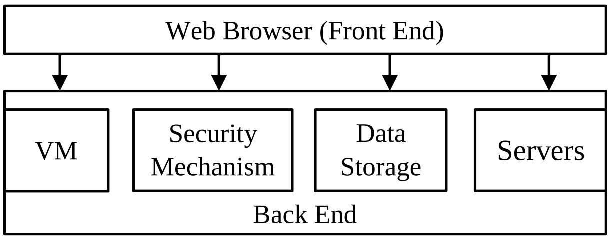

Fig.1. Cloud computing architecture

The way resources are demonstrated is a noteworthy time for the resources allocation to receive requests. The services a Cloud can offer designers can deliberate on many different levels and can be optimized with allocation in numerous parameters. At least these necessities for RAS to work appropriately must be considered when modelling and clarifying the resources. Cloud resources can be regarded as any physical or virtual resource designers can ask for from the cloud.

This request should be managed by the RAS in an adaptable and clear way. This adaptability should allow the physical resource to be utilized dynamically, thus avoiding the provision of resources beyond [6]. The cloud services in resources programming have a huge requirement as this schedule will also take the task schedule within the resources afterwards. A few examples of the single resource they can run at all times can be found. Accessibility and coherence must be checked and, moreover, the resources of a similar kind must contain a reasonable burden. The above parameters should have a diagnosis or function and the allocation should be done in an optimal way [7].

The major advantage of Cloud migration is the applications' scalability. The scalability of cloud resources, unlike grids that serve application requirements, allows the supply of resources in real time. Cloud computing offers less expensive services such as storage, computing, and bandwidth resources. User requirements are normally arranged for the tasks.

New scheduling techniques and methods must be planned in order to overcome the problems posed by network properties between users and resources. Some common planning methods can be used in new planning methods to integrate them into certain network mind methods in order to give answers for predetermined and more productive job planning. The tasks are usually arranged according to user needs [8].

2. RELATED WORK

This section examines the work of many other authors, who have come up with suggestions related to allocation of resources in the cloud computing environment.

In cloud computing, there are numerous techniques for assigning resources. In the Cloud computing environment, these network resource allocation strategies and the way in which these procedures can be executed are examined [10]. There is, however, an efficient resource allocation approach that meets all user requests and limits overall costs for users and other clouds. There are a number of planning algorithms for resource allocation.

The authors focused on two issues in the cloud computing environment [11]: planning and the allocation of resources. The Hadoop Map Reduce and its schedulers first depict the other problem in planning is to provide virtual machines with a cloud resource.

The infrastructure segment has a critical relationship to the use of power in a cloud environment [12]. Cloud groups with support of the cloud cluster node were proposed for energy use investigations. The combined energy use of the Cloud system is a mixture of energy usage nodes, all equipment used, storage energy use, power consumption by various parts such as fans, and the current conversion loss.

An enhanced FCFS resource algorithm is used to address the fundamental questions of cloud computing resource allocation planning. The customer server format depends on the architecture. This architecture includes an allocator of resources that identifies and reallocates customer resources [13].

With their respective parameters, the authors study various scheduling algorithms in different states. The schedule is carried out to achieve the greatest benefit and increase the competence of the workload. This is why we have distinctive kinds of algorithms for scheduling.

For the efficient distribution of resources in a cloud computing environment, priority-based allocation techniques are used without wasting resources. The algorithm requires batches of the user, e.g. the quantity of processor and user memory [14].

The Efficient Assignment of Resources approach reduces cloud providers' power consumption. In the cloud environment, energy consumption is one of the main concerns. Due to server operations and network communication, the problem is categorized into two main groups.

Virtualization is used for cloud resources. As networks, devices and various resources increase, heterogeneity also increases. In the cloud environment, the number of customers and resources are increasing. Consequently, investigating and finding the appropriate resources, due to scalability and complexity, is a difficult task. Various existing approaches have offered solutions for the allocation of resources. But, the Percentage of QoS needs to be efficiently increased. From the perspective of this requirement, this paper offers a solution based on a frame and an algorithm.

3. MULTI-STAGE FRAMEWORK WITH QOS BASED RESOURCE ALLOCATION USING RNN

The main goal of this paper is to find an economical solution for the efficient allocation of resources through an improved quality of service. By improving the performance of the entire system, customer satisfaction is increased. It can be achieved by virtualization of the server, which is a subpart of the allocation of resources. Using the server virtualization, resources, response times and costs are handled effectively. The work is based on future cloud computing technology in this paper. The main characteristics of the future Cloud computing are the geographical distribution of servers, low latency, localization and mobility.

Future Cloud computing is an extension of current cloud computing technologies, reducing the inconveniences of existing Cloud computing and increasing its efficiency. This article aims to increase the efficiency of the cloud computing environment in the future through an algorithm for resource allocation and its apps. For this purpose, the existing algorithms that are used to allocate resources are first analyzed and the framework that is developed for the proposed algorithm is then designed. The framework proposed for solving the problem of failure tolerance, overflow and low flow as set out in Fig.2 is developed. For the future cloud environment, a new model is designed as proposed to handle the aforementioned problem of resource allocation.

In this framework that is designed for the cloud environment in the future, three different layers of this framework are integrated into three different agents. The algorithm is initially deployed in a client-server and in the agent-side to meet the needs of the resources available to existing clients. All requests in the application queue are examined to determine the validity of the application. If the requests are not valid, they will be deleted and the resources will be checked. If there are no resources in the resource layer, the request is transferred to the server layer. The framework proposed is arranged in several phases. All requests, processor availability, request management, the finding of resources, and allocation of resources are monitored and investigated at each stage.

All data centers are arranged in the resource layer in a typical way. The more the data servers are connected to each Cloud communication in the Cloud environment, the more are the communication layers that are connected directly to the Cloud Data Center. Each server manager includes servers which verify the processor's accessibility and are responsible for virtual machinery management. All clients submit their requests to the manager of the server and load the requests to the manager of the server. According to the following procedure, the server management processes requests. If the sender does not receive a result in the specified time interval, the customer is assigned waiting status for the process. Clients are transmitted to server managers for the subsequent processes. The data server directly gives the customer the processor to improve the response time and sends a confirmation to the respective server manager.

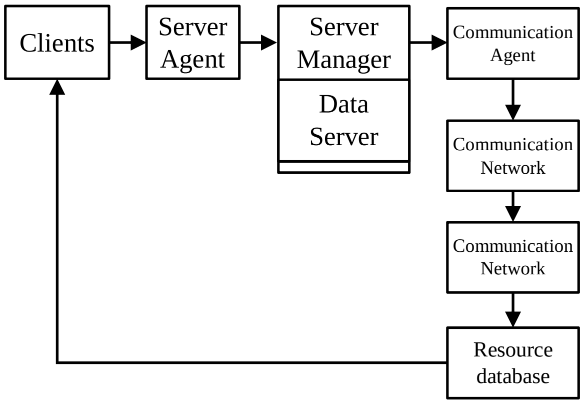

Fig.2. Proposed framework of QoS based resource allocation

4. PROPOSED RNN ALGORITHM

In the middle of each cloud stage, the aforementioned RNN algorithm is implemented and deployed. Each request is accepted in every middle layer and is not interrupted anywhere in the Cloud while processing all the requests from the client. If the application process is longer than the time assigned, the application shall be transmitted to the next middle levels. This is the reason why the RNN algorithm reduces the time and costs and the output of resources more effectively.

The Cloud environment is supposed to consist of servers of different types, many server managers, and many resources of varying types as determined by each server. The various servers are connected to heterogeneous data server resources. The resource bandwidth, memory, storage and CPU are computed each time and compared with the resource threshold values as per the allocation of calculated QoS parameters.

The energy of each layer device shall be examined during verification of request and response, validation and exchange from one layer to another layer. If all calculated resources QoS parameters fulfill the following limit, the resource is assigned. A power threshold value is assigned to all devices to exchange and transmit requests and answers. The REQ-RES is transmitted if the present energy of the devices exceeds or is equal to the energy threshold value, other energy appliances are chosen for transmission.

5. SIMULATION RESULTS AND DISCUSSION

In this section, the proposed RNN is deployed and simulated in the cloudsim tool. The results are calculated according to load, time, valid demand, allocated resources and energy. The results obtained are taken and compared to the current approach. It is apparent from the results, wherein values better than the existing values are obtained by the proposed RNN.

The time required to process the load also increases if the load increases. RNN requires 10ms for 40 MB of load processing, while ERA only requires 25 MB of load that is lower than the system proposed. The error during cloud computing RNN loading meets fewer errors, and ERA meets more errors. For example, for 250 requests RNN meets 0.1% of error but for that number ERA meets 0.3% of error. It is observed that RNN processes requests with very fewer errors compared to ERA.

In comparison with ERA, RNN also has a lower power consumption. For example, the cloud computing process RNN consumes 700 W for 200 requests while the ERA consumes 860 Watts of energy, and shows that the energy consumption of the RNN is efficient. Finally, the efficiency of resource allocation is indicated by frequency of load. The proportion of resource allocation increases as the load frequency generally.

The results obtained from the RNN algorithm simulation are represented by Fig.3 - Fig.6. The main aim here is to assess the RNN's QoS parameters. The time needed to process the requests is calculated and displayed in Fig.3 based on the load of the requests. This is done by changing the duration of the simulation time and calculating the load of requests. The time is increased by 10ms at each execution, from 10ms to 100ms. The RNN algorithm can handle 130 MB to 135 MB in processor loads from this simulation. The time taken by the RNN is also found to be considerably smaller than the ERA. Hence, RNN is time-efficient, whereas RNN is cost-effective, because the cost is only calculated on a time basis. Input requests and answers are initially validated for the allocation of resources. This calculates the error value to predict the validation of the request and the response. The lesser error percentage indicates that the response to the incoming requests is more accurate. The server manager processes the requests and assigns the answers only when the incoming requests and the existing answers are valid.

The simulation error is predicted and shown in Fig.4. In this paper, 500 jobs were assumed to be used to calculate the percentage of errors. Whereas, positive bug marks no resource for a valid request that can be found, and negative bug marks the request and the resource is not valid. The power efficiency of the entire framework is calculated to transfer the data after validating requests and answers. In this paper the energy efficiency and time delay is calculated and compared with the other threshold values based on the QoS factors prior to data processes. Therefore, the energy consumption of RNN is comparatively less than ERA.

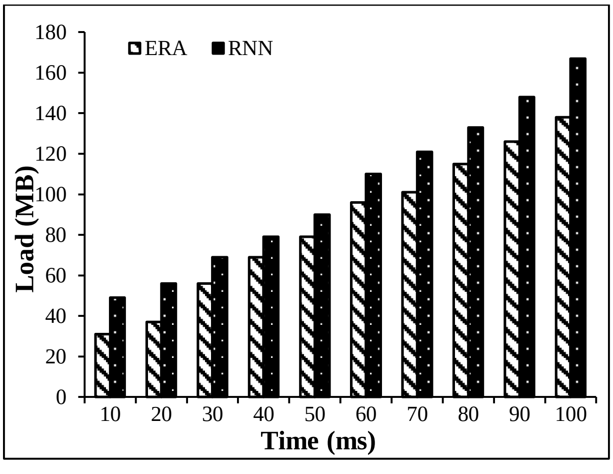

Fig.3. Time versus request's load performance comparison

The Fig.5 shows the comparison of energy consumption between the proposed RNN and ERA approach. From Fig.5, the proposed RNN is more important than the current ERA. Finally, a simulation is used to calculate the number of valid requests assigned with appropriate resources in terms of the valid requests. Input for load frequency at the same time, the efficiency of the resource allocation is calculated. The RNN based obtained result is compared with the existing results CA-DAG, CU-DAG and EB-DAG discussed in [17]. From the obtained result shown in Fig.6, it is noticed that the proposed RNN is better than the existing approaches it term of resource allocation.

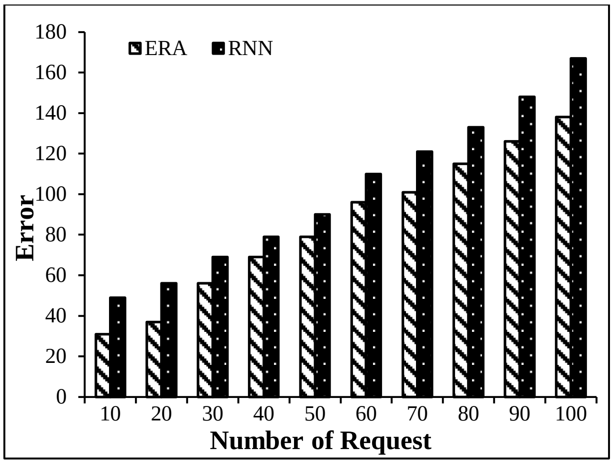

Fig.4. Estimation of average error

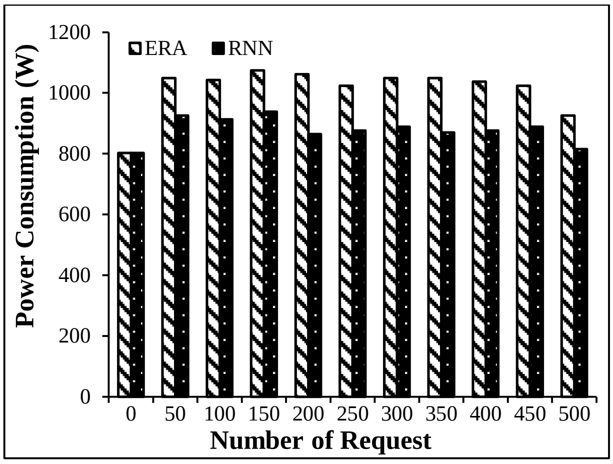

Fig.5. Comparison of power consumption

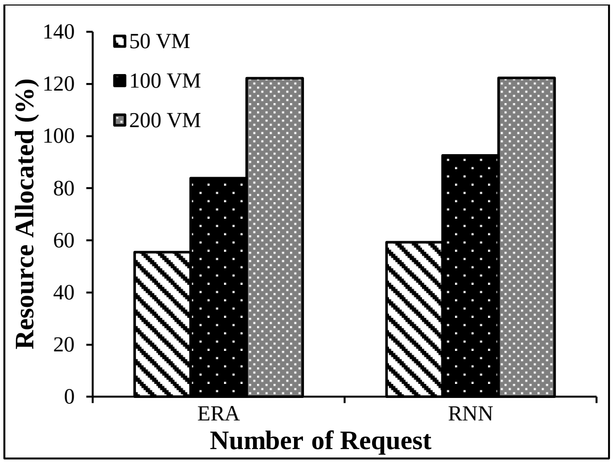

Fig.6. Number of requests vs. resource allocated

For an increased number of incoming applications, the number of valid applications is increased. The number of applications valid depends directly on the number of requests that the server manager receives. The QoS parameters such as time, cost, energy and the allocation of resources for the proposed RNN are efficient from these results and discussions as compared to the current ERA approach.

6. CONCLUSION

In order to enhance cloud computing efficiency, a multi-phase framework, RNN, has been proposed. The proposed framework reduces the amount of data required for exchange or passage towards processing and storage in the cloud. The proposed RNN method may transmit any IoT data to the Cloud, and RNN may thus provide server administrators with proximity to the client. It is very difficult to understand cloud computing without resource allocations because of scalability and infrastructure costs. From the result of the simulation, it is apparent that the proposed RNN Multi-Stage Framework is efficient. Compared to the ERA, an increase in efficiency for load handling and a reduction in energy consumption is demonstrated by the proposed RNN approach. In comparison to the CA-DAG approach, the RNN approach improves the allocation of resources.

REFERENCES

[1] Brian Underdahl, Margaret Lewis and Tim Mueting, "*Cloud Computing Clusters for Dummies*", Wiley Publication, 2010.

[2] V. Vinothina, R. Sridaran, and Padmavathi Ganapathi, "A Survey on Resource Allocation Strategies in Cloud Computing", *International Journal of Advanced Computer Science and Applications.*, Vol. 3, No. 6, pp. 97-104, 2012.

[3] Dzmitry Kliazovich, Johnatan E. Pecero, Andrei Tchernykh, Pascal Bouvry, Samee U. Khan and Albert Y. Zomaya, "CA-DAG:Modeling Communication-Aware Applications for Scheduling in Cloud Computing", *Journal of Grid Computing*, Vol. 12, No. 3, pp. 1-17, 2015.

[4] R. Sudeepa and H.S. Guruprasad, "Resource Allocation in Cloud Computing", *International Journal of Modern Communication Technologies and Research*, Vol. 2, No. 4, pp. 19-21, 2014.

[5] Shahin Vakilinia, Behdad Heidarpour and Mohamed Cheriet, "Energy Efficient Resource Allocation in Cloud Computing Environments", *IEEE Access*, Vol. 4, pp. 8544-8557, 2016.

[6] Shimpy , and Jagandeep Sidhu, "Different Scheduling Algorithms in Different Cloud Environment", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, No. 9, pp. 1-16, 2014.

[7] Parag R Kaveri, Vinay Chavan and Hemant Deshmukh, "A Study on Resource Oriented Computing", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.2, No. 6, pp. 25-28, 2012.

[8] N.R. Ram Mohan and E. Baburaj "Resource Allocation Techniques in Cloud Computing-Research Challenges for Applications", *Proceedings of IEEE International Conference on Computational Intelligence and Communication Networks*, pp. 556-560, 2012.

[9] M. Savani Nirav and Amar Buchade, "Priority Based Allocation in Cloud Computing", *International Journal of Engineering Research and Technology.*, Vol. 3, No. 5, pp. 111-115, 2014.

[10] T. Pering, T. Burd, and R. Brodersen, "The Simulation and Evaluation of Dynamic Voltage Scaling Algorithms", *Proceedings of IEEE International Symposium on Low Power Electronics and Design.*, pp. 76-81, 1998.

[11] Eman Elghoneimy, Othmane Bouhali and Hussein Alnuweiri, "Resource Allocation and Scheduling in Cloud Computing", *Proceedings of IEEE International Workshop on Computing, Networking and Communications.*, pp. 309-314, 2012.

[12] Liang Luo, Wenjun Wu, Dichen Di, Fei Zhang, Yizhou Yan and Yaokuan Mao, "A Resource Scheduling Algorithm of Cloud Computing based on Energy Efficient Optimization Methods", *Proceedings of the IEEE International Conference on Green Computing*, pp. 1-6, 2012.

[13] Sujit Tilak and Prof. Dipti Patil, "A Survey of Various Scheduling Algorithms in Cloud Environment", *International Journal of Engineering Inventions*, Vol. 1, No. 2, pp. 36-39, 2012.

[14] Aditya Marphatia, Aditi Muhnot, Tanveer Sachdeva, Esha Shukla and Lakshmi Kurup, "Optimization of FCFS Based Resource Provisioning Algorithm for Cloud Computing", *IOSR Journal of Computer Engineering*, Vol. 10, No. 5, pp. 1-5, 2013.