IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Survey on Resource Allocation Methods in Cloud Computing

Hilda Lawrance[*1], Dr.Salaja Silas[2]

[*1]Department of VLSI Design Bharath University, Chennai, India

[2]Assistant Professor – ECE Dept Bharath university, Chennai, India

Abstract

Cloud computing is a very popular area in current world which is rising very fast and the futures of the field seems really broad and strong. In order to provide quality of service in the cloud environment is a highly challenging task. The cloud clients should obtain reliable services from the provider based on their desire. In the particular cloud computing service, the resource allocation process is based on quality of service and cost of resource. The provider should allocate the resources in a proper way to render good services to the clients. This paper elucidates an elegant survey made on the different resource allocation methods used in the cloud computing environment.

Keywords: Cloud computing, Service Level Agreement, Resource Allocation Problem.

Introduction

Cloud computing is a very popular research area at present which has many challenges. Cloud environment makes it possible to access applications and associated data from anywhere.

Companies are able to rent resources from cloud for storage and other computational purposes so that their infrastructure cost can be reduced significantly. Further they can make use of company-wide access to applications, based on pay-as-you-go model [7]. It provides reliable, customized and QoS (Quality of Service) guaranteed computing dynamic environments for end-users.

The basic principle of cloud computing is that user data is not stored locally but is stored in the data center of internet. The companies which provide cloud computing service could manage and maintain the operation of these data centers. The users can access the stored data at any time by using Application Programming Interface (API) provided by cloud providers through any terminal equipment connected to the internet [7].

Cloud computing incorporates infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS). Infrastructure as a Service is a provision model in which an organization outsources the equipment used to support operations, including storage, hardware, firewalls, servers and networking components. The service provider owns the equipment and is responsible for housing, running and maintaining it. The client typically pays on a per-use basis. SaaS is a software delivery model in which software and its associated data are hosted centrally typically in the cloud and are typically accessed by users using a thin client, normally using a web browser over the Internet. SaaS has been incorporated into the strategy of all leading enterprise software companies. PaaS provides the entire infrastructure required to executes applications over the Internet. It is delivered in the same way as a utility like electricity or water. Users simply “tap in” and take what they need without worrying about the complexity behind the scenes. PaaS is based on a metering or subscription model so users only pay for what they use.

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Different customers needed different types of services and the services may vary while time changes. So resource allocation is an important task. Providers should allocate resources in a proper way to give good services to the clients. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module. By proper allocation of resources, the resource utility rate can be improved and service cost can be reduced. For the efficient utilization of resources, a good resource allocation strategy is needed.

There are different allocation models that are used in cloud computing area. Each of this models use certain methods and algorithms. Most of them are based on the cost and availability of the

resources. The following section elaborates on the methodologies of resource allocation models of the cloud computing environment. For the purpose of simplicity only the certain parts are surveyed in the vast cloud computing environment and the outcome of the survey is given below.

Different Resource Allocation Methods and Models

1) SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments **[1]**

SaaS is a software delivery method that provides access to software and its functions remotely as a Web-based service. It allows organizations to access business functionality at a cost typically less than paying for licensed applications since SaaS pricing is based on a monthly fee. In order to deliver hosted services to customers, SaaS companies have to either maintain their own hardware or rent it from infrastructure providers. This requirement means that SaaS providers will incur extra costs. Though the cost of the resources has to be minimum, it is also important to satisfy a minimum service level to customers. Saas providers are able to manage the variety of customers, mapping customer requests to infrastructure level parameters and considering heterogeneity of Virtual Machines.

The allocation method uses two different algorithms such as ProfminVmMaxAvaiSpace and ProfminVmMinAvaiSpace. First algorithm is designed to minimize the number of VMs by utilizing already initiated VMs. The criterion for reusing VM is, it should have maximum available space. The algorithm optimizes the profit by minimizing number of initiated VM. Moreover, it minimizes number of violations caused by service upgrade because VM has the maximum available space. In such a way, it reduces the penalty caused by upgrading service. The disadvantage of this algorithm is that it can decrease the profit. The maximum available space is occupied by small number of accounts and it leading other requests to be served by a new VM. To overcome the disadvantages of this algorithm, reducing the space wastage by using minimum available space (MinAvaiSpace) Strategy instead of MaxAvaiSpace Strategy. When there are more than one VM with same type, deployed with the same product type as customer request required, the VMs with enough available space to serve are selected. Then request is scheduled to the machine with the minimum available space in a best-fit manner). The proposed algorithms minimize the SaaS provider's cost and the number of SLA violations based on the dynamic allocation of resources to requests.

2) Dynamic resource allocation for spot markets in cloud computing environments **[2]**

IaaS means infrastructure as a Service. Users get different type of VMs (Virtual Machines) from the provider. This resource allocation approach is mainly implemented in the IaaS platform. The demands for virtual machines fluctuate frequently. The provider should allocate virtual machine with desired QoS to user in any situation. The IaaS provider adjusts the capacity of each VM for giving good services to users. The mentioned service provider should consider revenue and energy cost. If the demand for a particular IaaS service is very low, then the service provider should decrease the cost of service for attracting more customers. If same service demand is very high, then provider should increase the service charge for increasing income. Energy cost is another important factor. Power is necessary for working of servers in a datacenter. If unused servers consume power, then provider should pay extra money for unused servers. The provider should take measures to reduce energy for the unused servers. If the demand for a certain service provider is higher than the capacity of the datacenter, then the customer with highest priority gets the requested resources. This is dynamic capacity control problem. Model Predictive Control (MPC) is used for solving this problem. This approach gives the higher net income than the static allocation methods. The approach did not consider public cloud environment's demand response.

3) RAS-M: Resource Allocation Strategy based on Market Mechanism in Cloud Computing **[3]**

One of the most important objectives of cloud system is to provide high quality but transparent services to cloud clients through improving the utilization of the large Data Centers. However, the available resources of the providers and the resources requirement of consumers are both varied dynamically. To manage resources dynamically in terms of the varied requirements of consumers is a challenge in Cloud Computing environments. In order to manage this problem, a resource allocation strategy based on market mechanism (RAS-M) has been adapted. Architecture of RAS-M mainly consists of three parts: Consumer Agent (CA), Resource Agent (RA) and Market Economy Mechanism. Consumer Agent (CA) delegates the consumer to participate in the market system and aims to obtain maximal benefit for the consumer. RA delegates one type of resource to publish the resource's price and to adjust the price of according to the relationship of supply and demand in the market system. Market Economy Mechanism is responsible for balance the resource supply and the demand.

The method works as follows: Firstly, utility functions of CAs are constructed to denote their satisfaction with fractions allocated to them, then the equilibrium state is defined and its optimality is proved. Finally a GA-based price adjusted algorithm deals with the problem of balancing the demand and supply in market model. Under the Cloud Computing environment, RAS-M can improve the resource utilization while maximizing the benefits of all CAs. Both the consumers and providers of Cloud service gain their maximal profit by using this method. One of the disadvantages of this method is that, RAS-M is only implemented to allocate resource in the lowest level of Cloud Computing, and only manage the CPU resource.

4) Efficient Resource Scheduling in Data Centers using MRIS **[4]**

Most of the allocation algorithms have the drawback that, they typically consider either a one-dimensional resource type or treat all resources as one abstract capacity when determining the sequence or combination of applications. These resource modeling does not focus on the fact that applications need a variety of resources such as CPU power, memory, storage, network bandwidth and so on. Multi-dimensional Resource Integrated Scheduling (MRIS), is an algorithm to find optimal solution. This considers resource scheduling problem to a bounded multi-dimensional knapsack problem, taking into account the requirement dependency among multi-dimensional resources. A shared-resource pool prioritizes jobs differently by their preferred resource types. To meet the needs of resource requirement formalizing a job schedule model considering the dependency among multi-dimensional resources. The jobs with dissimilar demands also works and for scheduling multiple resources is proposed which will run effectively on a shared data center to achieve nearly optimal resource utilization in each dimension. The algorithm calculates the relationship between multi-dimensional resource supply and demand. Based on supply-demand ratio, the model assigns priority to jobs which require less scarce resources. The method has obtained high system efficiency and application performance. The disadvantage is that, it requires high processing time.

5) Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment **[5]**

Performance of the entire system is very important in cloud computing environment. A cloud computing service provider can increase the performance by allocating Virtual Machine (VM) to a suitable Physical Machine (PM). The service provider should utilize the physical memory resources effectively for reducing the cost. If the provider increases the utilization of the physical memory resources, and then the performance of the virtual machine inside physical memory is decreased. A physical memory can allocate more than one number of virtual machine. Because of this reason, the resource allocation on the cloud computing plays an important role. The provider needs to set a threshold level for PM utilization. The location is aware of dynamic resource allocation model which helps to improve the performance of the cloud service. The main two factors in this model are physical memory's location and physical memory's dynamic utilization. This model performs two main tasks. They are VM placement decision making and VM migration decision making. The provider performs the allocation of a particular VM to a suitable PM in VM placement decision making. If the provider gets a request from its user, then the specific provider finds the nearest PM in their area. The service provider finds an appropriate VM, this provider checks utilization level of that particular PM. If the PM's utilization level is applicable for allocating a virtual machine, then the provider allocates the virtual machine to that particular PM. The cloud computing provider monitors each physical memory regularly. VM migration decision making is mainly for performing migration in virtual machines. The provider performs the VM migration, when the utilization level of a particular physical memory is more than the threshold level. All physical machines in a particular datacenter send their utilization level requests to the provider. The provider finds the suitable PM for allocating the virtual machine. The provider allocates the VM from one physical memory to another physical memory. The advantage of this model is that can prevent the degradation of the data center performance. This model is not implemented in a real cloud computing environment. This model has not deal about the performance in real cloud computing environment.

6) Cloud Simulation Resource Allocation Algorithm Based on Multi-Dimension Quality of Service (CS-RSA algorithm) **[6]**

Most of the resource allocation algorithms only consider about single Quality of Service (QoS) parameters. For the proper allocation and for the best utilization of resources, it is necessary to consider multiple QoS parameters such as CPU speed, bandwidth, stability, length of tasks and so on. This method takes into consideration of multi dimensional QoS parameters. It considers values of all quality of service (QoS) parameters for every task and creates a task matrix. In the same way resource QoS parameters are taking and obtaining resource matrix. By using Analytic Hierarchy Process (AHP), calculates the weight of QoS parameters. Compute synthetic need volume of each task and arranging in

big to small order in accordance with the task synthetic QoS need volume and taking first task from the task line. Computing user satisfaction of that particular task with every resource and finding out the resource which gives highest satisfaction. If that particular resource is not in use, allocate that resource to task. Delete the particular task from task line. Check whether the task line is blank, if not repeat the procedures for allocation. CS-RSA algorithm have several advantages. It considers multi QoS parameters for allocation since most of the allocation methods are taking only single parameter. It reduces task finishing time for allocation and greatly improves resource utility rate. It also meets customer satisfaction. Since the proposed method is fully utilizing resource capability and operating difference of various resources is small, therefore, loading equilibrium of algorithm is high.

Conclusion

Cloud computing is the next generation of technology which unifies everything into one. It is an on demand service because it offers dynamic flexible resource allocation for reliable and guaranteed services in pay as-you-use manner to public. The review shows that dynamic resource allocation is growing need of cloud providers for more number of users and with the less response time. An effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper discusses different resource allocation strategies based on the QoS parameter and dynamic nature cloud. Some of the algorithms discussed above will consider only single type of QoS parameter which will decrease efficiency of the method. CS-RSA algorithm considers user satisfaction along with multi QoS parameters. The resource utility rate and efficiency of the algorithm is high while comparing to other allocation methods. This will also helps to decrease task finishing time.

References

[1] Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya, 2011. “SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments”. 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Pages 195-204.

[2] Qi Zhang, Eren G¨urses, Raouf Boutaba, Jin Xiao, 2011. “Dynamic Resource Allocation for Spot Markets in Clouds”. Fourth IEEE International Conference on Utility and Cloud Computing (UCC), Pages:178-185.

[3] Xindong YOU, Xianghua XU, Jian Wan, Dongjin YU, 2009. “RAS-M: Resource Allocation Strategy based on Market Mechanism in Cloud Computing” Fourth China Grid Annual Conference. DOI: 10.1109/ChinaGrid.2009.41.

[4] Nagendram S., Vijaya Lakshmi J., Venkata Narasimha Rao D., Naga Jyothi Ch, 2011. “Efficient Resource Scheduling in Data Centers using MRIS” Indian Journal of Computer Science and Engineering (IJCSE), Vol.2, Issue 5, pages: 764-769.

[5] Gihun Jung, Kwang Mong Sim, 2011. “Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment”, International Conference on Information and Computer Applications, Dubai, 2011.

[6] Wuqi Gao and Fengju Kang, 2012. “Cloud Simulation Resource Scheduling Algorithm Based on Multi-Dimension Quality Of Service”. Information Technology Journal, 11: 94-101.

[7] V. Vinothina, Dr.R.Sridaran, Dr.PadmavathiGanapathi., 2012. “A Survey on Resource Allocation Strategies in Cloud Computing” International Journal of Advanced Computer Science and Applications, Vol. 3, No.6.