A SURVEY ON VARIOUS RESOURCE ALLOCATION POLICIES IN CLOUD COMPUTING ENVIRONMENT

Vaghela Ankita

PG student, Department of Computer Engineering, Alpha College of Engineering and Technology, Gujarat, India, ***ankita.solanki27@gmail.com***

Abstract

Cloud computing is bringing a revolution in computing environment replacing traditional software installations, licensing issues into complete on-demand services through internet. In Cloud computing multiple cloud users can request number of cloud services simultaneously. So there must be a provision that all resources are made available to requesting user in efficient manner to satisfy their need. Resource allocation is based on quality of service and service level agreement. In cloud computing environment, to allocate resources to the user there are several methods but provider should consider the efficient way to guarantee that the applications' requirements are attended to correctly and satisfy the user's need This paper survey different resource allocation policies used in cloud computing environment.

Keywords: *Cloud computing, Resource allocation*

---***--

1. INTRODUCTION

Cloud Computing is a technology that uses the internet and central remote servers to maintain data and applications. Cloud computing allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access. This technology allows for much more efficient computing by centralizing data storage, processing and bandwidth.

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources e.g., networks, servers, storage, applications, and services that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Cloud computing customers do not own the physical infrastructure; rather they rent the usage from a third party provider. They consume resources as a service and pay only for resources that they use.

Cloud computing provide three types of services[7], including software as a service (SaaS),platform as a service (PaaS) and infrastructure as a service (IaaS). In Software as a service consumers purchase the ability to access and use an application or service that is hosted in the cloud. A benchmark example of this is Salesforce.com, as discussed previously, where necessary information for the interaction between the consumer and the service is hosted as part of the service in the cloud. In Platform as a service Consumers purchase access to the platforms, enabling them to deploy their own software and applications in the cloud. The operating systems and network access are not managed by the consumer, and there might be constraints as to which applications can be deployed. In Infrastructure as a service Consumers control and manage the systems in terms of the operating systems, applications, storage, and network connectivity, but do not themselves control the cloud infrastructure.

Resource Allocation

Resource allocation [8] is a subject that has been addressed in many computing areas, such as operating systems, grid computing and datacenter management. A Resource Allocation System (RAS) in Cloud Computing can be seen as any mechanism that aims to guarantee that the applications' requirements are attended to correctly by the provider's infrastructure. Along with this guarantee to the developer, resource allocation mechanisms should also consider the current status of each resource in the Cloud environment, in order to apply algorithms to better allocate physical and/or virtual resources to developers' applications, thus minimizing the operational cost of the cloud environment.

Cloud resources can be seen as any resource (physical or virtual) that developers may request from the Cloud. For example, developers can have network requirements, such as bandwidth and delay, and computational requirements, such as CPU, memory and storage. Generally, resources are located in a datacenter that is shared by multiple clients, and should be dynamically assigned and adjusted according to demand. It is important to note that the clients and developers may see those finite resources as unlimited and the tool that will make this possible is the RAS. The RAS should deal with these unpredictable requests in an elastic and transparent way. This elasticity should allow the dynamic use of physical resources, thus avoiding both the under-provisioning and over-provisioning of resources.

There are different resource allocation policies that are used in cloud computing environment. Each of this policy uses certain methods and algorithms which are given below:

2. DIFFERENT RESOURCE ALLOCATION POLICIES

A time-driven adaptive mechanism for cloud resource allocation [1]

Cloud computing service providers deliver their resources based on virtualization to satisfy the demands of users. In cloud computing, the amount of resources required can vary per user request. Therefore, the providers have to offer different amounts of virtualized resources per request. To provide worldwide service, a provider may have data centers that are geographically distributed throughout the world. Likewise, the user locations vary in geographic location. Since cloud computing services are delivered over the internet, there may be undesirable response latency between the users and the data centers. Hence, for the best service, the provider needs to find a data center and physical machine that has a light workload and is geographically close to the user. The proposed model finds the best match for the user requests based on two evaluations: 1) the geographical distances between the user and data centers and 2) the workload of data centers. Hence, the model allows the users to find a data center that is guaranteed to be the closest distance and have the lightest workload. Also, it finds a light workload physical machine within the data center for a provider.

Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems [2]

Resource allocation is the most important challenges in cloud computing. The service provider should work hard for allocating resources based on the client's SLA (Service Level Agreement).Force directed search algorithm is the solution for SLA based resource allocation problem for multi-tier applications in cloud computing .This algorithm considers the Gold SLA, and Bronze SLA. The provider gives the guarantee for the response time in Gold SLA. The requests are moved forward and backward in multi-tier service model. The server serves the backward requests. Probability Distribution Function (PDF) is used for finding the arrival rate in the Gold SLA. The resource management problem's aim is to maximize the total profit. The profit maximization problem has these steps. It performs the summation of client's utilities. It calculates the operation cost of the services. This problem does the multiplication the total power consumed during the decision making time. It selects only one tier for servers. It finds the ON servers. It calculates the allocated memory for clients. The profit maximum problem is solved using upper bound on the total profit. The force between clients and servers are the major factor in the resource consolidation using force directed search algorithm. This algorithm takes a client based on the highest force towards a new server. It performs the load replacement, if the server is available. The algorithm performs the updating of forces between the clients and servers. If there is no positive force differential, then the algorithm stops its working. The algorithm saves the best solution in each step.

Multi-dimensional Resource Allocation Algorithm in cloud Computing [3]

Cloud computing has emerged as a new technology and it has been increasingly adopted in many areas including science and engineering as well as business. How to arrange large-scale jobs submitted to cloud in order to optimize resource allocation and reduce cost is an issue of common concern. Paper present are two common ways to optimize resource utilization. One is at the application level when applications are arriving, other is in the period of applications running. In this paper, author makes effort on the former way to address multi-dimensional resource allocation problem by proposing a resource allocation scheme using fewer nodes to process user's applications. To address multi-dimensional resource allocation problem, raises several concerns. Firstly, allocate method should decide which virtual machines should be assigned with a new set of jobs. Secondly, there exists an optimal set of nodes which can process new arriving applications, how to find an efficient way to assign applications in nodes is another issue should be solved. In response to these issues, we use virtual machine as the minimum resource allocation unit. When a new batch of applications arrives, applications is decompose into several types of jobs, each job with the same type has the same requirement of resource. Aiming at the first issue mentioned above, author formulate it by adding multi-dimensional constraints in resource allocation process, which assure jobs can be processed in nodes selected by object function. Aiming at the second problem, under the purpose of optimize utility of nodes, this nodes select problem as a binary integer programming problem, and our object function can assure using working nodes' remain resources to process more jobs. Unlike the existing resource allocation schemes which allocate more physical resources (CPU, memory, etc.) to the exiting virtual machine, this model assign jobs in running nodes to make nodes work at a higher utilization.

SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments [4]

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 Prof min Vm Max AvaiSpace and ProfminVmMinAvaiSpace. First algorithm is designed to minimize the number of VMs by utilizing already initiated VMs. The criterion for reusing VMis, 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 hasthe 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.

Adaptive Resource Allocation for Pre-empt able Jobs in Cloud Systems [5]

In this paper authors propose an adaptive resource allocation algorithm for the cloud system with preempt able tasks in which algorithms adjust the resource allocation adaptively based on the updated of the actual task executions. Adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS) algorithms are use for task scheduling which includes static task scheduling, for static resource allocation, is generated offline. The online adaptive procedure is use for re-evaluating the remaining static resource allocation repeatedly with predefined frequency. In each reevaluation process, the schedulers are re-calculating the finish time of their respective submitted tasks, not the tasks that are assign to that cloud.

Policy based resource allocation in IaaS cloud [6]Most of the Infrastructure as a Service (IaaS) clouds use simple resource allocation policies like immediate and best effort. Immediate allocation policy allocates there sources if available, otherwise the request is rejected. Best-effort policy also allocates the requested resources if available otherwise the request is placed in a FIFO queue. It is not possible for a cloud provider to satisfy all the requests due to finite resources at a time. Haizea is are source lease manager that tries to address these issues by introducing complex resource allocation policies. Haizea uses resource leases as resource allocation abstraction and implements these leases by allocating Virtual Machines (VMs). Haizea supports four kinds of resource allocation policies: immediate, best effort, advanced reservation and deadline sensitive. Proposed dynamic planning based scheduling algorithm is implemented in Haizea that can admit new leases and prepare the schedule whenever a new lease can be accommodated. Experiments results show that it maximizes resource utilization and acceptance of leases compared to the existing algorithm of Haizea.

3. COMPARISON

Technique	Method	Parameter	Findings
Time-Driven adaptive mechanismfor cloud resource allocation	Finding work load of data center and distance between user and data center	Response time	Better response time and resource utilization
Multi-Dimensional SLA-based resource allocationfor Multi-tier Cloud Computing Systems	Force directed search algorithm	Service Level Agreement	Maximize total profit
A Multi-Dimensional resource allocation Algorithm in cloud **Computing**	Multi-Dimensional Resource allocation algorithm	Use lightest node to allocate resource	Increase Resource Utilization and reduce cost of data center

SLA-based resource allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments	ProfminVmMaxAvaiSpace and ProfminVmMinAvaiSpace algorithm	Service Level Agreement and Cost	Minimize the SaaS provider's cost and the number of SLA violations
Adaptive Resource Allocation for Pre-empt able Jobs in Cloud Systems	Adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS) algorithms	Service Level Agreement	Increase Resource Utilization
Policy based resource allocation in IaaS cloud	Four policies used Immediate, best effort, advanced reservation and deadline sensitive	Service Level Agreement	Maximize Resource Utilization

CONCLUSION

Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software and information are provided to users over the network. It is anon demand service because it offers dynamic flexible resource allocation for reliable and guaranteed services in pay as-you-use manner to public. This paper discusses various resource allocation policies based on service level agreement and quality of service. All above methods used different techniques for allocating resource to the user. Time driven gives better response time and increase resource utilization. Multidimensional SLA based algorithm save resources and increase resource utilization. Multidimensional algorithm increase resource utilization and reduce cost of data center. SLA-based policy minimizes the sass provider's cost and the number of SLA violations. Adaptive resource allocation policy increase resource utilization. Policy based resource allocation maximize resource utilization.

REFERENCES

[1] GIHUN Jung; KWANG MONG Sim; PAUL C. K. Kwok; MINJIE Zhang. A TIME-DRIVEN Adaptive Mechanism for Cloud Resource Allocation. Proceedings of IC-BNMT2011, 2011, PP441-446

[2] HadiGoudarzi, MassoudPedram, Multi-dimensional SLAbased Resource Allocation for Multi-tier Cloud Computing Systems, in Proceedings of IEEE International Conference on Cloud Computing (CLOUD),Washington DC USA, 2011.

[3] Bo Yin, Ying Wang, LuomingMeng, XuesongQiu, A Multi-dimensional Resource Allocation Algorithm in Cloud Computing , Beijing University of Posts and Telecommunications, Beijing 100876, China, 2012

[4] Linlin Wu, Saurabh Kumar Garg andRajkumarBuyya, 2011. "SLA-basedResource Allocation for Software as aService Provider (SaaS) in CloudComputing Environments". 11thIEEE/ACM International Symposium onCluster, Cloud and Grid Computing, Pages195-204.

[5] Jiayin Li, MeikangQiu, Jian-Wei Niu, Yu Chen, Zhong Ming, "Adaptive Resource Allocation for Pre-empt able Jobs in Cloud Systems," in 10th International Conference on Intelligent System Design and Application, Jan. 2011, pp. 31-36.

[6] AmitNathani, Sanjay Chaudharya, GauravSomani, "Policy based resource allocation inIaaS cloud", Future Generation Computer Systems 28 (2012) 94–103 doi: 10.1016/ j.future. / 2011.05.016

[7] Introduction to Cloud Computing, White paper, Dialogic, Making Innovation Thrive.

[8] GlaucoEstácioGonçalves, PatríciaTakako Endo, ThiagoDamasceno, Resource allocation in clouds: Concepts, Tools and Research Challenges.

BIOGRAPHIES

Vaghela Ankita (M.E), AlphaCollege of Engineering and Technology, Khatraj, Kalol