International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 1, January - 2014

A Survey on Dynamic Resource Allocation Techniques in Cloud Computing

Vineet Khisty
PG Student of Information and Technology, Department of Computer Science and Technology, Karunya university, Coimbatore, India.

Mrs. Jeba Priya.
Assistant Professor, Department of Computer Science and Technology, Karunya university, Coimbatore, India.

Abstract

Various researches have been carried out in the field of Cloud computing to find out solutions for various challenges faced by Cloud. Due to the ever increasing demands of the users for services or resources, it becomes difficult to allocate resources accurately to the user demands in order to satisfy their requests and also to take care of the Service Level Agreements (SLA) provided by the service providers. This paper discusses a brief discussion of three Dynamic Resource Allocation techniques namely, Location-aware Dynamic Resource Allocation Model in Cloud Computing Environment, Priority Based Dynamic Resource Allocation in Cloud Computing and Automated Negotiation with Decommitment for Resource allocation in Cloud Computing.

1. Introduction

Cloud computing is an emerging technology and a new way of computing. Cloud computing has a unique advantage where the service providers can provide the services and the clients can utilize the services . Service providers may provide various types of services and also provide certain Service Level Agreements (SLAs) to the users which have to be fulfilled without making any compromisation. These services provided by the service providers should be made available to the users as and when they are being demanded. Cloud computing uses the concept of virtualization to meet the ever increasing demands from the users. Virtualization is used since the physical resources available are very limited to meet the demands of the users. Thus, virtualization helps in creating multiple virtual instances of a single physical resource and these virtual instances can be independently be given to the individual requests of the users. These virtual instances created from a single physical resource will have all the abilities of the physical resources.

There are several challenges which are being faced by cloud computing. Some of which are security, service delivery and billing, interoperability, reliability, performance and bandwidth cost. Along with these challenges mentioned, there is one more major concern that cloud has to deal with. This major concern is Resource management. Resource management helps the providers of cloud services to efficiently manage the scarce resources available and to accurately allocate these resources to the demands of the users in order to improve the utilization and also to avoid performance degradation of the system. Efficient management of resources helps to improve the utilization of the system and also helps in improving the performance. Resource management helps the service providers to provide the Service Level Agreements (SLAs) to the users without making any compromises.

This paper consists of Six sections. Section 2 of this paper deals with the related work which gives a brief introduction of various resource allocation techniques. Section 3 deals with the dynamic resource allocation strategies in which Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment, Priority Based Dynamic resource allocation in Cloud Computing and Automated Negotiation with Decommitment for Dynamic Resource Allocation in Cloud Computing techniques are discussed in detail. Section 4 provides the advantages and disadvantages of Dynamic Resource Allocation Techniques. Section 5 provides the Conclusion of the paper. Section 6 provides the references.

2. Related Works.

Accurately allocating the resources dynamically to the user requests is a very challenging problem in the resource management. It is considered to be the most critical challenge in cloud computing, since allocation of the resources to the user requests should be done in such a way that there should be no performance degradation and that it will increase the utilization of the system. To overcome this issue, different researchers have come up with various techniques. Some of the techniques will be explained in brief in this paper. This section will provide introduction to some of the Dynamic resource allocation techniques.

In [4] the authors have compared various dynamic resource allocation techniques. In [5], the authors of the paper have proposed an architecture called as Topology Aware Resource Allocation which uses a prediction engine with a lightweight simulator to estimate the performance of a given resource allocation and a genetic algorithm to find an optimized solution in large search space. In [6] authors have proposed algorithm to allocate the resources using linear scheduling strategies. In [7], the authors have proposed an architecture known as the Nephele framework to dynamically allocate the resources for parallel data processing.

The following section deals with the brief explanation of dynamic resources allocation strategies. These techniques which would be briefly discussed in the following section are Location-aware Dynamic Resource allocation model in cloud computing environment, Priority Based Dynamic resource allocation in cloud computing and Automated Negotiation with Decommitment for Dynamic resource allocation in cloud computing.

3. Dynamic resource allocation strategies in cloud computing

This paper does not deal or suggest any particular algorithms or Architecture and also does not propose any new Dynamic resource allocation strategies. It is just a survey paper to provide an overview over some of the dynamic resource allocation techniques presently available in cloud computing environment. This section provides brief working of some of the Dynamic Resource allocation techniques that are being proposed by various researchers from all over the world.

There are 3 techniques that have been briefly described in this section. They are:

- Location –Aware Dynamic Resource Allocation Model for Cloud Computing Environment.
- Priority Based Dynamic resource allocation in Cloud Computing.
- Automated Negotiation with Decommitment for Dynamic Resource Allocation in Cloud Computing.

3.1. Location-aware dynamic resource allocation model for cloud computing environment.

In [1], the authors have proposed a new architecture which makes the decision of placing the virtual machines on the physical machines based on the location of the users who have requested for resources and also it considers the location of the data centres. It allocates the virtual machine to that physical machine whose utilization level is proper and that it would not affect the performance of other allocated virtual machines. This technique of placing the virtual machine to the physical machine based on the location of the user will decreases the response time, since the placement of the virtual machine will be done to that physical machine which will be placed close to the user location. The proposed idea is demonstrated using an Agent-Based cloud computing system.

3.1.1. Agent-based cloud computing system.

The proposed scheme is based on the Agent-Based Cloud Computing System. The proposed system allows the user to send requests to access an arbitrary value of resources from anywhere and at any time. In the above figure, the system contains two components in the data centre. They are: 1) Data centre super nodes (DCSN), 2) sets of Physical Machines (PMs).

a. Data Center Super Node (DCSN)

The function of Data Centre Super Node (DCSN) is to keep a record of the resource utilization reports of all the physical machines present in the data centre. It also has a function to make decision for allocating the Virtual Machines to the appropriate physical machines based on the location and its utilization level. To carry out these functions the DCSN consists of two subcomponents: 1) Report Repository which stores the physical machines current utilization level reports, 2) Decision Making Engine (DME) which enables to take decisions on placing the new Virtual Machines to the appropriate physical machines while they are still in the running mode based on the current reports present in the Report Repository.

b. Sets of physical machines (PM)

The physical machines are nothing but real resources that can accommodate many virtual machine instances. Hypervisor is a specialized virtualization layer present in the physical machines (PMs). This hypervisor has a specialized virtual machine (VM) known as Domain zero (DOM 0) as shown in the figure. This Dom 0 consists of a resource monitoring agent that will monitor the resource utilization. There is also a User Monitoring Agent (UMA) that will report the utilization level of the running application to the Machine Monitoring Agent (MMA).

3.2. Priority based dynamic resource allocation in cloud computing.

In [2], the authors have proposed a

dynamic resource allocation technique which is based on the priority of the tasks. The authors have proposed an algorithm which pre-empts the execution of the current task having low priority with the task having a higher priority. If both the current task and the next arrived task have the same priority and it is unable to pre-empt the current task, then the algorithm proposes to create a new virtual machine from the available resources. In this paper the authors have used certain *Use Techniques* in order to develop their proposed priority based dynamic resource allocation algorithm. The following are the *Use Techniques* that are being considered

a. Load Balancer.

The authors of this paper have utilized the load balancer algorithm. In this algorithm, the first step is to provide a list of the appropriate virtual machines as an input along with the customers' service deployment requests (R) and the application data (A). Then it gets the available physical resources (AR) and also the appropriate virtual machine (AP) which can be sufficient to provide the requested services. The next step is to collect the number of available virtual machines from various cloud and provide a list. Then select an appropriate virtual machine in order to satisfy the requested service from the list of the available virtual machines. After selecting the virtual machine, it is added to the list of used virtual machines so that the load balancer would not use the same virtual machine in the next iteration.

b. Forming a list of tasks based on the priorities.

Once the scheduler receives the users' requests, it partitions these requests based on the priorities. Here the initial static allocation takes place. In this paper, the authors have mentioned two greedy algorithms in order to generate the static allocation namely, the cloud list scheduling (CLS) and the cloud min-min scheduling (CMM). Here in this technique a list of task is formed which is based on priorities. Thus, after receiving the list of all the tasks which are being arranged by their priorities, the allocation of the resources to the task can be done.

3.3. Automated negotiation with decommitment for dynamic resource allocation in cloud computing.

In [3], the authors of the paper describe a Negotiation Protocol in which there are two entities namely, Buyer and seller. The Buyer entity refers to the users who requests for

resources and Seller are the service provider who provide the requested resources to the buyers. In between these two entities there is also an agent entity. This agent entity makes certain contracts in order to bind a set of resources from a provider to the consumer for a fixed time interval. Here the agent entity can also decommit from a contract by paying certain penalty to the other contract party.

3.3.1. Negotiation protocol.

The initial state is known as the "*buyer reasoning*" . In this state, the buyer '*b*' decides the way in which the offer has to be made. After the buyer '*b*' decides about the offer and sends it to the sender '*s*', the state changes to "*Seller reasoning*". In this state itself seller decides whether to accept or reject the offer sent by the buyer. If s accepts the offer, a tentative agreement is made. Otherwise, s sends a bid to b and then it is b's turn to decide its offer**.** If the tentative agreement is confirmed by the buyer '*b*' the negotiation is in the "*final agreement*" state. If the tentative agreement is not accepted by one of the agent, then the whole negotiation between the sender and the buyer will be lost and the buyer '*b'* can again make an offer. Along with the decommit action , there are other two actions introduced by the authors of this paper, *"confirm"* and *"cancel"*. With these actions, whenever a seller accepts the offer, the buyer can still have the chance to *"decommit"* from the agreement with no payment of penalty. Assume if the buyer '*b'* needs only a single resource . In absence of the *"cancel"*, if buyer *'b'* sends an offer to multiple number of sellers and all accept, b must buy multiple items or decommit from agreements by paying penalties . Due to the presence of actions *"confirm"and "cancel",* buyer *'b' can* choose only one contract while negotiating with multiple sellers simultaneously.

4. Advantages and limitations of dynamic resource allocation techniques.

In [4], the authors of the paper have discussed the advantages and limitations of Dynamic Resource Allocation. They are:

Advantage:

- There is no hardware or software overhead
- Reduced Location overhead. The data and applications can be accessed from anywhere.
- Resource sharing amongst cloud providers can be done during lack of resource supply.

Limitations:

- As the resources are rented by the users from the remote servers there is a lack of control over their resources.
- Migration problem occurs, when the user wants to switch to some other provider for the better storage of their data
- Spread of malware is easy due to interconnection of the servers on cloud.

5. Conclusion.

Cloud computing is an emerging technology and being various researches have been carried out in order to solve the challenges faced by cloud. There are several challenges that cloud is facing, out of which a major challenge being the dynamic resource allocation techniques are discussed in this paper. This paper provides an overview of three Dynamic resource allocation techniques presently being found out. Along with the three techniques some more techniques are being introduced in this paper. Hence the main motive behind this survey paper is not to propose any

new idea or concept in the resource allocation field, but a simple survey is being made in this field in order to motivate future researches to be carried in this area and eventually provide a solution to completely overcome this cloud computing challenge and strengthen the cloud computing paradigm.

6. References.

[1]. Gihun Jung and Kwang Mong Sim, *Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment*, International Conference on Information and Computer Applications (ICICA), IACSIT Press, Singapore, 2012.

[2].Chandrashekhar S. Pawar and Rajnikant B. Wagh, *Priority Based Dynamic Resource Allocation in Cloud Computing* Cloud and Services Computing (ISCOS), 2012. International Symposium .

[3]. Bo An, Victor Lesser, David Irwin and Michael Zink, *Automated Negotiation with Decommitment for Dynamic Resource Allocation in Cloud Computing*, Conference at University of Massachusetts, Amherst, USA.

[4]. Ronak Patel and Sanjay Patel, *Survey on Resource Allocation Strategies in Cloud Computing*, International Journal of Engineering Research and Technology (IJERT), Vol.2 Issue 2 , February 2013.

[5]. Gunho Lee, Niraj Tolia, Parthasarathy Ranganathan, and Randy H.Katz, *Topology aware resource allocation for data intensive workloads*, ACM SIGCOMM Computer Communication Review, 41(1):120—124, 2011.

[6]. Abirami S.P. and Shalini Ramanathan, *Linear scheduling strategy for resource allocation in cloud environment*, International Journal on Cloud Computing: Services and Architecture(IJCCSA),2(1):9 17,2012.

[7]. Daniel Warneke and Odej Kao, *Exploiting dynamic resource allocation for efficient parallel data processing in the cloud*, IEEE Transactions on Parallel And Distributed Systems, 2011.