ISSN 2278-733X

SURVEY: RESOURCE MANAGEMENT STRATEGIES IN CLOUD ENVIRONMENT

Rasmi.K [1], Vivek.V [2]

Department of Computer Science,karunya university,Coimbatore-641 114,Tamilnadu

Abstract **- Cloud computing is an on demand technology as it offers dynamic and supple resource allocation for trustworthy and guaranteed services in pay as-you-use way to public. The speciality of this technology is that the any number of cloud services can be concurrently accessed by any number of users. So it is indispensable that every user should get the sufficient resources in a proficient manner. The resource allocation in cloud computing is nothing but integrating the cloud provider deeds in order to exploit and allocate scarce resources. The service level agreement satisfaction is very significant concerning the user as well as the service provider. Bare minimum SLA violation brings utmost customer satisfaction. Here in this paper a survey is carried out on the area of resource allocation strategies which tries to preserves the customer satisfaction to its maximum. The merits and demerits of each technique are also discussed.**

Keywords **– cloud computing, QoS, service level agreements, resource overbooking, resource management strategies.**

I. INTRODUCTION

Cloud computing accommodate the infrastructure, software and platform as a services. And it emerges as a new computing paradigm which aims to provide consistent, customized and QoS (Quality of Service) affirmed computing dynamic environments for the end customers. Today the capital issues accompanying to cloud computing are the minimum data transfer cost, network bandwidth and less response time in data transfer. The basic principle of cloud computing is that user data is not stored locally and it is stored in the data centre of internet.

There are a number of advantages for the cloud computing technique is that it possesses lower costs services, re-provisioning of resources and remote accessibility. Cloud computing lowers the cost by avoiding capital expenditure by the company in renting the physical infrastructure from a third party provider.

In cloud computing the resource allocation has got a cogent role in the performance of the whole process and the level of customer satisfaction provided by the process. But while providing the maximum customer satisfaction the service provider ought to be definite the profits that incur to them also. So the resource allocation ought to be efficient on both perspectives i.e. on the end user and the service provider point of view.

In order to get such a process the new technologies insist that the process ought to be with maximum SLA (Service Level Agreements) violation. The service level agreement is an element of the terms that is provided by the service provider to give assurance to the end user about the level of service that it can provide to the end user. In short, for a customer high QoS means few SLA violations. The remainder of the paper is organized as follows: section II explains a number of the resource allocation strategies that are existing today and by which way they are giving maximum customer satisfaction or efficient resource allocation. In section III it talks about the merits and demerits of the existing strategies. The idea of overbooking and its advantages are discussed under section IV. Finally conclusions are drawn in section V.

II. RESOURCE MANAGEMENT STRATEGIES

Cloud computing that is rooted in resources acquired on demand is generating a lot of interest among service providers and consumers. Here in this section we are going to analyse the resource allocation and reallocation (load balancing) methods that are previously present in the cloud environment and their basic principle.

A. Autonomic Workload Provisioning (AWP)

Executing applications in the underlying resources has got two key steps in the case of cloud environment. The first step is called the VM provisioning. It consists of creating VM instances with the intent of hosting each application request and then the harmonizing of the specific features and the needs of the request. And the second step in this process is mapping and forecasting these requests onto distributed physical resources. That is known as resource provisioning. Most of the virtualized data centers currently provide a set of general purpose VM classes and they are provided with generic resource configurations. The speciality is that it quickly becomes insufficient to support the highly varied workloads.

In [1] it proposes a new autonomic workload provisioning that addresses the challenges of enterprise grids and clouds. The main aim of this mechanism is that to improve the resource utilization and which is achieved with the help of reducing the overprovisioning. It can be reached through two levels.

In order to reduce the overprovisioning caused by the difference between the virtual resources allocated to VM instances and those requested by the individual jobs a new mechanism is introduced. And this technique is base on decentralized online clustering, and it helps to characterize the resource requirement classes and it is used for proactive VM provisioning.

This paper also introduced another way of resolving the overprovisioning problem which may be happened due to inaccuracies in client resource requests. This paper also explored the use of workload modelling techniques and their application.

The mechanism for dynamic and decentralized VM provisioning monitors the flow of arriving jobs from different queues in a decentralized manner during ongoing analysis windows of duration in the order of the startup time of new VMs.

The decentralized clustering mechanism possesses several advantages. It has the capability of analysing jobs across a dynamic set of distributed queues. It has no prior knowledge regarding the existence of the number of clustering classes and also it has the adaptively changing behaviour towards the dynamic workloads and resources.

The main drawback of this method is that it is not suitable for real-time application. It is because of the fact that the resource provisioning is done after the queuing of the requests. That is each request has to wait in the multiple queues before they get serviced. This causes some reduction in the QoS factor and the user satisfaction.

B. Linear Scheduling Strategy

The resource allocation is taken into consideration commonly the parameters like CPU utilization, memory utilization and throughput etc. The cloud environment has to take into reflection all these things for each of its clients and could provide maximum service to all of its clients. In [2] it suggests that when we are taking the scheduling of resources and tasks separately it imposes large waiting time and response time. In order to overcome this shortcoming this paper introduces a new approach namely Linear Scheduling for Tasks and Resources (LSTR).

Here scheduling algorithms primarily focus on the distribution of the resources along with the requestors which will make best use of the selected QoS parameters. The QoS parameter selected in this approach is the cost function. The scheduling algorithm is designed based on the tasks and the available virtual machines together and specified as LSTR scheduling strategy. This is designed so as to maximize the resource utilization.

Scheduling algorithm is carried out based on the prediction that the initial response to the request is made only after collecting the resource for a finite amount of time but not allocating the resource as they arrive. But dynamic allocation could be carried out by the scheduler dynamically on request for extra resources. This is obtained by

the continuous evaluation of the threshold value within the system.

Here the authors state that this approach is appropriate when we are considering the "shortest job first" rather than the FCFS approach. It is for the reason that the algorithm sorts the requests by excluding the arrival times. It only considers the "threshold" of the request for the scheduling purpose.

This approach has the advantage that it has an improved throughput and response time. But it is not appropriate for the interactive real-time areas because there is no consideration for the arrival time. For interactive applications the requests are treated in a "first come first serve" basis.

C. Pre-Copy Approach

Clark et al. [3], talks about the live migration of the virtual machines. In this paper they suggest that migration of the operating system instances across distinct physical hosts is a useful tool for the administrator of data centers and clusters. It also provides a separation between hardware and software and facilitates load balancing, fault management and low level system maintenance.

In "Pre-Copy Approach" pages of memory are iteratively copied from the source machine to the destination host and in addition there is a fact that all these things are done without ever stopping the execution of the system. Pagelevel protection hardware is used to make sure that a consistent snapshot is transferred. For controlling the traffic of other running services a rate-adaptive algorithm is used. And during the final phase it pauses the virtual machine and copies any remaining pages to the destination and after that resumes the execution there.

Franco et al. [4] put forward some of the drawbacks that are encountered in the above mentioned approach. It points out that the conventional approach in [3] is inadequate because of the high RTTs and potential store and forward handling of virtual machines. It also points out that it will result in long forwarding chains. This will create a delay to the user experiences with the system.

D. MIPS Based Scheduling

One of the important requirements for a cloud computing environment is to provide reliable QoS. It can be able to define in terms of Service Level Agreements, it describes about such characteristics like maximal response time, minimal throughput, or latency delivered by the system.

In [5] it talks about a system comprises of a large-scale Cloud data center comprising of diverse physical nodes. Here each node has a CPU, which can be a multicore and the speciality is that its performance is defined in Millions Instructions Per Second (MIPS). The speciality of the MIPS is that it depends deeply on the instructions to be executed [6]. The usual way to measure CPU potential is FLOPS (Floating Point Operations per Second). And in FLOPS it will specify which type of instruction you are dealing with. In short "MIPS" cannot be a well-founded way to judge processors for their performance.

E. Match Making and Scheduling

Shikharesh et al. [8], suggests that the "Match making" is the first step and "scheduling" is second in the resource allocation in cloud environment. Matchmaking is the procedure for allocating jobs associated with user requests to resources selected from the available resource pool. Scheduling refers to determining the order in which jobs mapped to a specific resource are to be executed [8]. It also points out that there are some uncertainties that are associated with such type of "match making" and scheduling. They can be like

- Error Associated with Estimation of Job Execution Times

It is considered that estimating the execution time for a line of work is a very hard labour and faults may happen very often. There is one deviant condition known as the formation of "resource idle time". It is bechanced because of certain unwanted statuses like jobs may run for a smaller time in comparison to their estimated execution time. There is one more reason is that abnormal ending of the jobs. These give rise to a serious debasement in system performance because jobs that could have used the resource throughout these idle time span might have been turned away by the matchmaker that anticipated the resource to be committed executing the job with an over estimated execution

time. There is one other extreme for the estimation of the job execution time. That is happened when the estimated time is less than the actual execution time. And the under estimation of job execution times may lead to job terminations because the resource may be booked for executing another job right after the completion of the first job's execution. Both of the above conditions i.e. over estimation and under estimation of job execution time are undesirable.

- Lack of Knowledge of Local Resource Management Policies

Matchmaking is challenging in cloud systems because the scheduling policy used at every resource may not be known to the resource broker. Resource broker performs admission control for advanced reservations during the request to resource mapping. This negative condition happens because of the fact that the exact system configuration for a cloud may not be fully known during the time of system design or deployment. After all it may change many times during the lifetime of the entire system.

So the method given in [8] is vulnerable to some sort of uncertainties that are explained above.

F. Just-In-Time Resource Allocation

In [10] it talks about the cost based workload provisioning and "just- in- time resource allocation".

- Workload Prediction

Here the prediction of the workload on the application and estimation of the system behavior over the prediction horizon is using a performance model. Here optimization of the system behavior is carried on by taking into consideration the minimization of the cost incurred to the application. This cost can be a combination of various factors such as cost of SLA violations, leasing cost of resources and a cost associated with the changes to the configuration. The advantage of such types of methods is that it can be applied over various performance management problems from systems with simple linear dynamics to systems with complex dynamics. The performance model can also be varied and affected with system dynamics as conditions in the environments like workload variation or errors in the system change.

- Just-in-time Resource Allocation

To optimize resource usage and to minimize the number of idle resources, an ideal solution is to set a time interval and change resources as many times according to workload changes. Within the limit of this interval resources are changed continuously in accordance with the change in load, assuming we can always over estimate the load. The limit of the interval is made too small. This extreme will make ensure that the optimum number of resources is always being used. Clearly, such as scheme is not possible since changing resources is not spontaneous. And it also makes some problems in the cost related aspects.

In this just in time resource allocation the three components of the cost function refer individually to the penalty for violation of SLA bounds, cost of leasing a machine, and cost of reconfiguring the application when machines are either leased or released.

But for the look-ahead implementation of the time interval for each task need the implementation of recursive data structures. And the prediction of this look-ahead time also results in some prediction error [10].

III. COMPARISON

Here in this section it carries out a brief comparison between the resource management strategies discussed above. The merits and demerits of each method is mentioned. Table 3.1 gives the overall summary of the comparisons made.

IV. RESOURCE OVERBOOKING

Resource overbooking is nothing but reserving resources in advance. In [9] it gives a detailed description of the overbooking technique and what are the advantages that the customer can benefit from this technique in a cloud. It is more useful in the concept of virtualization, clearly say virtual desktops.

Resource overbooking is the technique that can establish an increase of the average utilization of hosts in a data center by reserving fewer resources than required in worst case. Since more virtual desktops can be allocated to a host, the cost for the service provider related to investment in hardware equipment, server maintenance cost and energy cost can be reduced.

TABLE I.COMPARISON BETWEEN THE RESOURCE MANAGEMENT STRATEGIES

Author	Method	Merits	Demerits
Quiroz et al. [1]	AutonomicWorkload Provisioning	Maximum resource utilization, reduced overprovisioning.	Queing of requests, not suitable for real-time applications.
Abirami S.P.,Shalini Ramanathan [2]	LinearScheduling Strategy	Improved throughput and response time.	Not suitable for interactive real time applications
Clark et al. [3]	Precopy Approach	Page level protection hardware	Long forwarding chains, delayed user experiences.
Anton Beloglazov ,Rajkumar Buyya [5]	MIPS based reallocation	Depends on instructions to be executed.	Not the proper way for measuring the CPU performance.
Shikharesh Mujumdar [8]	Match making and scheduling	Cost effective, less delay	Uncertainties that are associated with such type of "match making" . Error Associated with Estimation of Job Execution Times. Lack of Knowledge of Local Resource Management Policies
Roy et al. [10]	Just-in-time Resource allocation	Cost effective	Prediction error and use of recursive data structures.

The risk parameter that is limiting the degree of overbooking is the risk to affect the user satisfaction [11]. Fiedler in this proposed method suggests a careful overbooking for network virtualization and it also obeys service level agreements (SLA) for full and limited availability. Full availability means the availability of all the required resources. Limited availability stands for the availability of certain share of required resources that are statically guaranteed at given degrees.

Urgaonkar et al. in [12] mentioned about how to maximize the revenue through overbooking. They suggests that provisioning cluster resources based on the worst case needs of the application results in low average utilization. it is because of the fact that average resource requirements of an application are normally smaller than its worst case requirements. And also the resources tend to idle when at times when the application does not utilize its peak reserved share. In [12] it summarizes that in shared hosting platforms techniques to overbook (i.e. under-provision) resources in a guarded manner will outcome in revenue maximization through optimized usage.

V. CONCLUSION

Nowadays cloud computing technology is increasingly being used in enterprises and business markets. In cloud environments, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper summarizes different resource management strategies and its impacts in

cloud system. Here also mentioned the concept of overbooking and its advantages and certain limiting factors. And it is found that the overbooking concept has got a close relationship with effective resource management in cloud environment.

References

[1] Quiroz A, Kim H, Parashar M, Gnanasambandam N, Sharma N; "Towards workload provisioning for enterprise grids and clouds". 10th IEEE/ACM international conference on grid computing, 2009. pp 50-57.

[2] Abirami S.P. ,Shalini Ramanathan; "Linear Scheduling Strategy for Resource allocation in Cloud Environment"; International Journal on Cloud Computing and Architecture ,vol.2, No.1, February 2012.

[3] Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hanseny, Eric July, Christian Limpach, Ian Pratt, Andrew Warfield; "Live Migration of Virtual Machines", 2nd Symposium on Networked Systems Design and Implementation (NSDI) , May 2005.

[4] Franco Travostino, Paul Daspit, Leon Gommans, Chetan Jog, Cees de Laat, Joe Mambretti, Inder Monga,Bas van Oudenaarde, Satish Raghunath, Phil Wang; "Seamless Live Migration of Virtual Machines over the MAN/WAN" ; Elsevier Future Generation Computer Systems 2006.

[5] Anton Beloglazov ,Rajkumar Buyya ; "Energy Efficient Resource Management in Virtualized Cloud Data Centers"; 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing 2010.

[6] http://www.tomshardware.com/forum/285409-28-mips

[7] Stephen S. Yau , Ho G. An. "Adaptive Resource Allocation for Service-Based Systems", International Journal of Software and Informatics ISSN 1673-7288, Vol.3, No.4, December 2009, pp. 483–499.

[8] ShikhareshMujumdar; "Resource management on cloud :Handling uncertainities in parameters and policies" CSI communications , 2011 , edn. pp.16-19.

[9] Lien Deboosere , Bert Vankeirsbilck ,Pieter Simoens , Filip De Turck , Bart Dhoedt and Piet Demeester, "Efficient resource management for virtual desktop cloud computing", Springer 2012.

[10] Nilabja Roy, Abhishek Dubey and Aniruddha Gokhale ; "Efficient Autoscaling in the Cloud using Predictive Models for Workload Forecasting".

[11] Markus Fiedler, "On Resource Sharing and Careful Overbooking for Network Virtualization", 20th ITC Special Seminar, May 2009.

[12] Bhuvan Urgaonkar, Prashant Shenoy and Timothy Roscoe, "Resource Overbooking and Application Profiling in Shared Hosting Platforms", ACM Trans Internet Tecnology2009