International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 2 No. 4 Apr. 2013, pp. 525-532

http://www.issr-journals.org/ijias/

Resource Management Techniques in Cloud Environment - A Brief Survey

K. Rasmi and V. Vivek

Department of Computer Science and Engineering, Karunya University,
Coimbatore, Tamilnadu, India-641 114

ABSTRACT: Cloud computing referred to as the on demand technology because it offers dynamic and versatile resource allocation for reliable and warranted services in pay as-you-use manner to public. It is a technology that uses the web and central remote servers to take care of data and applications and permits users to use applications without installation and access their personal files at any computer with the assistance of internet access. This technology allows rather more efficient computing by consolidative data storage, processing and bandwidth. The specialty of this technology is that any variety of cloud services can be simultaneously accessed by any variety of users. So it is necessary that every user should get sufficient resources in a well-organized manner. The resource allocation in cloud computing is nothing but integrating the cloud provider activities in order to utilize and allocate scarce resources. The service level agreement satisfaction is incredibly necessary concerning the user as well as the service provider. Minimum SLA violation brings most client satisfaction. Here in this paper a survey is meted out on the realm of resource management strategies that tries to preserve the customer satisfaction to its maximum. There are some metrics which are able to evaluate the potency of these resource allocation strategies. The deserves and demerits of each technique are also mentioned.

KEYWORDS: Cloud computing, Service Level Agreements (SLA) violation, load balancing, QoS.

1 INTRODUCTION

Cloud computing provides the infrastructure, software and platform as a services. And it emerges as a brand new computing paradigm that aims to supply reliable, custom-made and QoS (Quality of Service) warranted computing dynamic environments for the end customers. The main problems related to cloud computing are the network bandwidth, response time, minimum delay in data transfer and minimum transfer cost for data. The basic principle of cloud computing is that user data is not stored locally and it is stored in the data centre of internet. The basic cloud design is shown in figure 1.

NIST Definition of Cloud Computing: "Cloud computing is a model for enabling 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" [19].

There are a variety of benefits for the cloud computing technique that it possesses lower cost services, re-provisioning of resources and remote accessibility. Cloud computing lowers the cost by avoiding capital expenditure by the corporate in dealing the physical infrastructure from a third party provider.

In cloud computing the resource allocation possesses an awfully vital role in the performance of the entire system and also the level of customer satisfaction provided by the system. However while providing the utmost customer satisfaction the service provider ought to make sure the profits that incur to them also. So the resource allocation should be economical on both views i.e. on the end user and the service provider perspective. So as to get such a system the new technologies insist that the system should be with minimum SLA (Service Level Agreements) violation. The service level agreement is a part of

Corresponding Author: K. Rasmi (rasmi2815@gmail.com)

the terms that is offered by the service provider to give assurance to the end user regarding the level of service that it can provide to the end user. In short, for a customer high QoS suggests few SLA violations [9].

Cloud computing has its signature in each aspects of life. For instance in today's business world with the amount of economic worsening and loss happening every day, the requirement for reliable, affordable technology is needed more than ever. Cloud computing is in a position to fill that void. Cloud computing offers its customer reliable service at versatile costs [17]. Educational clouds [18] have become extremely popular in recent years. It will create some vital modifications to the traditional educational system and the ways which were adopted by the later method. The remainder of the paper is organized as follows: section 2 explains some of the resource management strategies that exist today and by which way they are giving maximum customer satisfaction or efficient resource allocation. In section 3 it talks about the merits and demerits of the prevailing strategies. In section 4 analyzes the assorted parameters that are affecting the efficiency of resource management strategies and a comparison of the matrices are also made. Finally conclusions are drawn in section 5.

2 RESOURCE MANAGEMENT STRATEGIES

Cloud computing that is based on resources acquired on demand is generating a good deal of interest among service providers and consumers. Here during this section we are aiming to analyze the resource allocation and reallocation (load balancing) methods that are already present in the cloud environment and their bedrocks.

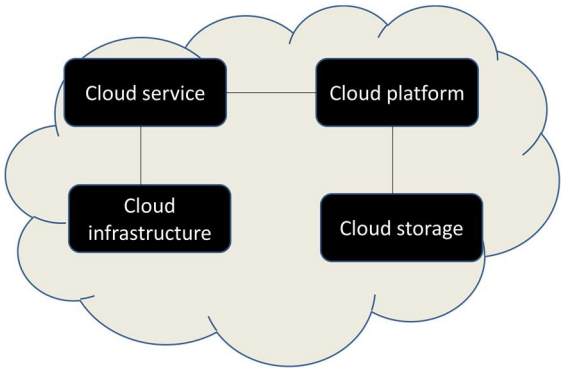

Fig. 1. Basic cloud architecture

2.1 LINEAR SCHEDULING STRATEGY

The resource allocation is taken into thought usually the parameters like CPU utilization, memory utilization and throughput etc. The cloud environment has got to take into consideration of these things for every of its clients and could offer maximum service to all of them. In [2] it suggests that when we are taking the scheduling of resources and tasks in an individual basis it imposes giant waiting time and response time. So as to beat this drawback a new approach specifically Linear Scheduling for Tasks and Resources (LSTR) is introduced. Here scheduling algorithms mainly target on the distribution of the resources among the requestors which is able to maximize the chosen 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 named LSTR scheduling strategy. This is often designed so as to maximize the resource utilization.

The scheduling algorithm is meted out based on the prediction that the initial response to the request is formed solely when assembling the resource for a finite amount of time (say 1 day or 1 hr like that) but not allocating the resource as they arrive. However dynamic allocation could be carried out by the scheduler dynamically on request for a few extra resources. This is often achieved by the continuous evaluation of the threshold value in the system. The authors states that this approach suitable when we consider the "shortest job first (SJF)" instead of the "first come first serve (FCFS)" way of scheduling. The algorithm sorts the requests by excluding the arrival times. It solely considers the "threshold" of the request for the scheduling purpose.

This approach has the advantage that it has a better throughput and response time. The only disadvantage is that it is not appropriate for the interactive real-time applications because it doesn't take into consideration the arrival time. For

interactive real time applications the requests are considered in a “first come first serve” manner. So the arrival time is important regarding this type of systems.

2.2 PRE-COPY APPROACH FOR SCHEDULING

In [3], it talks regarding the live migration of the virtual machines. Clark et al. suggest that migration of the operating system instances across distinct physical hosts is a great tool for the administrator of data centers and clusters. It in addition offers a separation between hardware and software and provides fault management, low level system maintenance and load balancing. Clark et al. came out with the idea of “pre-copy approach”. In this approach pages of memory are repeatedly copied from the source machine to the destination host and additionally there is an undeniable fact that all these things are done without ever stopping the execution of the system. Pagelevel protection hardware is employed to make sure that a consistent snapshot is transferred. For controlling the traffic of different running services a rate-adaptive algorithm is used. And during the final phase it pauses the virtual machine and copies any leftover pages to the destination and afterwards resumes the execution there. The factors touching the total migration are link bandwidth, migration overhead and page dirtied rate [20].

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

2.3 MATCH MAKING AND SCHEDULING

In [8] it tells that the “Match making” is the first step and “scheduling” is second within the resource allocation in cloud environment. Matchmaking is that the method of allocating jobs associated with user requests to resources designated from the obtainable resource pool. Scheduling refers to determining the order in which jobs mapped to a selected resource are to be executed [8]. It additionally tells that there are some uncertainties that are associated with such type of “match making” and scheduling. They can be like

2.3.1 ERROR ASSOCIATED WITH ESTIMATION OF JOB EXECUTION TIMES

It is considered that estimating the execution time for a job is a terribly laborious task and errors might happen fairly often. There is one abnormal condition referred to as the formation of “resource idle time”. It is happened because of certain unwanted conditions like jobs may run for a smaller time compared to their estimated execution time. There is one more reason known as abnormal termination of the jobs. These give rise to a serious degradation in system performance because jobs that could have used the resource during these idle time periods might have been turned away by the matchmaker that expected the resource to be busy executing the job with an over estimated execution time. Other problem raises the estimated time is less than the actual execution time. The under estimation of job execution times might lead to job terminations because the resource may be booked for executing another job right after the completion of the primary job’s execution. Both of the above conditions i.e. over estimation and under estimation of job execution time are unattractive.

2.3.2 LACK OF KNOWLEDGE REGARDING LOCAL RESOURCE MANAGEMENT POLICIES

We know that a cloud is a large and heterogeneous environment that may encompass variety of resources and each of them administered by their own local scheduling policies. Therefore matchmaking is tricky in such a system because the scheduling policy used at each resource may not be known to the resource broker. Resource broker performs admission control for advanced reservations at some stages in the request to resource mapping. This off-putting 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.

2.4 JUST-IN-TIME RESOURCE ALLOCATION

Roy et al. [10] illustrates about the cost based workload provisioning and "just- in- time resource allocation".

Workload Prediction is that the prediction of the workload on the application and assessment of the system behavior over the prediction horizon by employing a performance model. Here optimization of the system behavior is dispensed by taking into concern the step-down of the cost incurred to the application. This cost can be a mixture of varied factors like cost of SLA violations, cost associated with the changes to the configuration, and leasing cost of resources. The advantage of such kinds of strategies is that it can be applied over various performance management issues from systems with easy linear dynamics to systems with complicated dynamics. The performance model may also be varied and affected with system dynamics as conditions within the environments like workload variation or errors in the system modification.

2.4.1 JUST-IN-TIME RESOURCE ALLOCATION (JITRA)

To optimize resource usage and to reduce the number of idle resources, a perfect solution is to set a time interval and alter resources as persistently keeping with workload changes. Within the limit of this interval resources are changed unceasingly in accordance with the modification in load, assuming we are able to continually overestimate the load. The limit of the interval is made too tiny. 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 JitRA 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 it needs the implementation of recursive data structures. And the prediction of this look-ahead time also results in some prediction error [10].

2.5 MIYAKODORI: A MECHANISM FOR MEMORY REUSE

MiyakoDori [14], is a memory reusing mechanism to reduce the amount of transferred data in a live migrating system. When we are considering the case of dynamic VM consolidation, virtual machines may migrate back to the host where it was once executed and so the memory image in that host can be reused, thus contributing to shorter migration time and greater optimizations by VM placement algorithms.

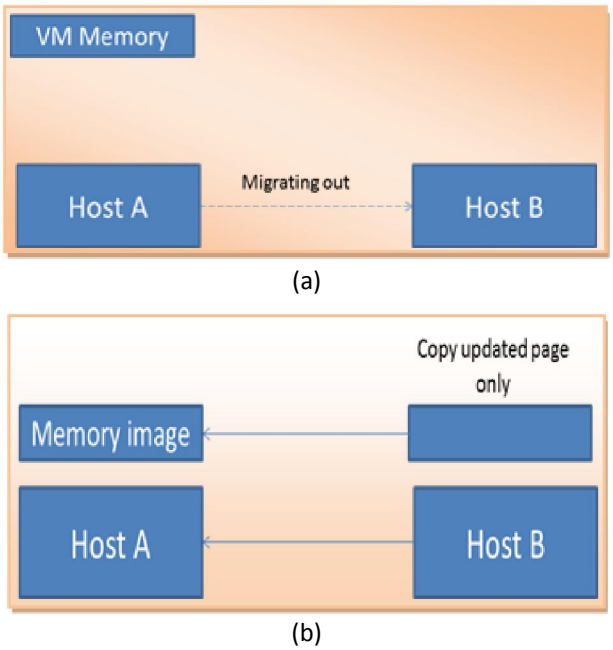

Fig. 2. (a) migrating out from host A to host B. (b) process of memory reusing while migrating back to the same host

In [15] it shows that this technique enables to reduce the total migration time. In this technique dirty pages alone need to be transferred to the former host.

2.6 ENHANCED LOAD BALANCING ALGORITHM (ELBA)

According to the design in [16] various users submit their diverse applications to the cloud service provider through the communication channel. There the Cloud Manager in the cloud service provider's datacenter being the prime entity to distribute the execution load among all the VMs by keeping track of the status of the VM. Cloud Manager maintains a data structure containing the VM ID, Job ID of the jobs that has to be allocated to the corresponding VM and VM Status to keep track of the load distribution. The VM Status represents the percentage of utilization. After that Cloud Manager allocates the resources and distributes the load as per the data structure. The Cloud Manager analyzes the VM status routinely to distribute the execution load in an equal manner. In course of processing, if any VM is overloaded then the jobs are migrated to the VM which is underutilized by tracking the data structure. If there are more than one available VM then assignment will be based on the least hop time. By the time of completion of execution, the Cloud Manager automatically updates the data structure.

The aim of load balancing in the cloud computing environment is to provide on demand resources with high availability. But often load balancing approaches suffer from various overheads. And they also fail to avoid deadlocks when there more requests competing for the same resource at the same time when the available resources are insufficient to service the arrived requests. The overall system layout is shown in figure 3.

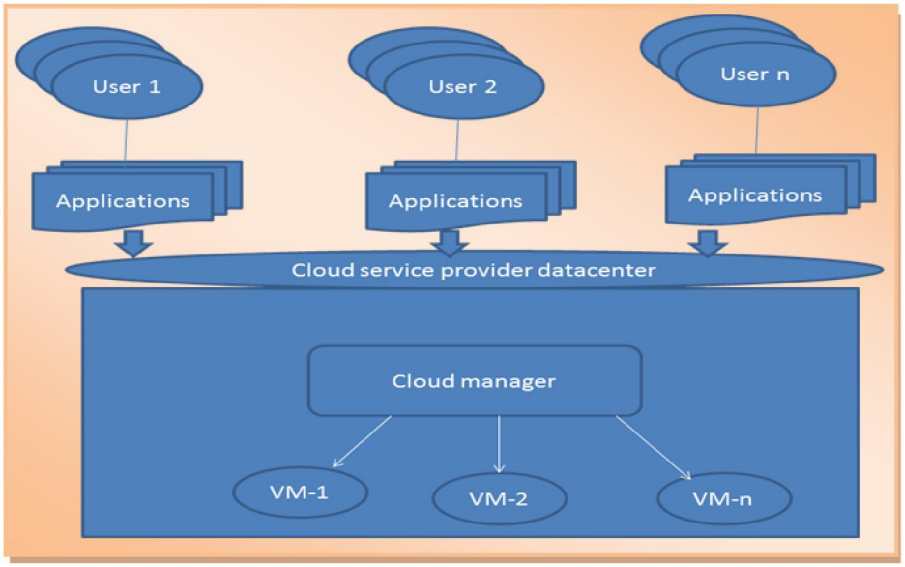

Fig. 3. Logical diagram for load balancer in ELBA

The ELBA approach using the efficient cloud management system helps to overcome the aforementioned limitations. This approach yields less response time compared to the existing approach. Less response time reduces job rejections and accelerates the business performance.

3 PROS AND CONS OF RESOURCE MANAGEMENT STRATEGIES

Here in this section it carries out a brief assessment between the resource management strategies discussed above. The qualities and demerits of each method are mentioned. Table 1 gives the overall summary of the comparisons made.

Table 1. Comparison between the resource management strategies

Author	Method	Merits	Demerits
Abirami S.P., Shalini Ramanathan [2]	Linear Scheduling Strategy	1. Improved throughput 2. Response time. 3. Improved resource utilization.	1.Not suitable for interactive real time applications
Clark et al. [3]	Pre-copy Approach	1.Page level protection hardware	1. Long forwarding chains. 2. Delayed user experiences.
ShikhareshMujumdar[8]	Match making and scheduling	1.Cost effective 2. Less delay	1. Uncertainties that are associated with such type of "match making". 2. Error Associated with Estimation of Job Execution Times. 3. Lack of Knowledge regarding Local Resource Management Policies
Roy et al. [10]	Just-in-time Resource allocation	1.Cost effective	1.Prediction error 2. Use of recursive data structures.
Akiyama et al.[14]	MiyakoDori	1. Memory reuse 2. Shorter migration time	1. Efficient only in cases where migration back to the same system.
Rashmi et al. [16]	Enhanced Load Balancing Approach	1. Less response time 2. High performance 3. Avoids deadlock 4. No overheads	1. Not cost effective

4 METRICS ASSOCIATED WITH RESOURCE MANAGEMENT STRATEGIES

Resource management is achieved through some sort of load balancing among the participating nodes. There are some metrics that will help to evaluate the efficiency of each load balancing method. Load balancing techniques in cloud environment, consider various parameters like performance, scalability, response time, throughput, resource utilization, fault tolerance, migration time and associated overhead [13].

- **Overhead Associated** - determines the amount of overhead involved while implementing load balancing algorithms. It is composed of overhead due to movement of tasks, inter-process and inter-processor communication. This metric should be minimized so that a load balancing technique can work efficiently.
- **Throughput –**it is used to calculate the number of tasks whose execution has been completed. It should be high to improve the performance of the system.
- **Performance –** is used to check the efficiency of the system. It has to be improved at a reasonable cost e.g. reduce response time while keeping some acceptable delays.
- **Resource Utilization –**it is used to check the utilization of resources in a system. It should be optimized for an efficient load balancing.
- **Scalability** - is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.
- **Response Time** - is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. It should be minimal to improve efficiency.
- **Fault Tolerance** - is the ability of an algorithm to perform uniform load balancing in spite of arbitrary node or link failure. Every system is expected to be highly fault tolerant.
- **Migration time** - is the time to migrate the jobs or resources from one node to other. It should be minimized in order to enhance the performance of the system. Or we can say that minimum migration is preferred by every efficient system.
- **Energy Consumption** - determines the energy consumption of all the resources in the system. Load balancing helps in avoiding overheating by balancing the workload across all the nodes of a Cloud, hence reducing energy consumption.

We can also make some comparisons based on the above matrices to evaluate or analyze the effectiveness of each of the resource management strategies. Table 2 gives the efficiency of each of the strategies based on some predefined matrices.

Table 2. RAS (Resource Allocation Strategies) efficiency analysis based on various matrices

Resource management strategy	Overhead associated	Throughput	Performance	Resource Utilization	Scalability	Response time	Fault tolerance	Migration time	Energy consumption
Linear Scheduling Strategy	✓	X	X	✓	✓	X	✓	✓	X
Pre-Copy Approach	✓	✓	X	✓	✓	X	✓	X	X
Match making and scheduling	✓	✓	X	✓	✓	X	X	✓	X
Just-in-time Resource allocation	✓	✓	✓	✓	✓	✓	X	X	X
MiyakoDori	✓	X	✓	✓	✓	✓	✓	X	X
ELBA	X	✓	✓	✓	✓	✓	✓	✓	X

5 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. It tries to analyze the resource allocation strategies based on various matrices and it points out that some of the strategies are efficient than others in some aspects. So the usability of each of the method can be varied from application area. That is one strategy which is suitable for real time interactive application may not be suitable for some other application area.

REFERENCES

[1] Rajkamal Kaur Grewal, Pushpendra Kumar Pateriya (2012), "A Rule-based Approach for Effective Resource Provisioning in Hybrid Cloud Environment", International Journal of Computer Science and Informatics ISSN: 2231 –5292, Vol-1, Issue-4.

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

[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 (2006), "Seamless Live Migration of Virtual Machines over the MAN/WAN", Elsevier Future Generation Computer Systems.

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

[6] Tom's Hardware [Online] Available: http://www.tomshardware.com/forum/285409-28-mips (2013)

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

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

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

[10] Nilabja Roy, Abhishek Dubey and Aniruddha Gokhale, "Efficient Autoscaling in the Cloud using Predictive Models for Workload Forecasting".
[11] Markus Fiedler (2009), "On Resource Sharing and Careful Overbooking for Network Virtualization", 20th ITC Special Seminar, May.
[12] Bhuvan Urgaonkar, Prashant Shenoy and Timothy Roscoe (2009), "Resource Overbooking and Application Profiling in Shared Hosting Platforms", ACM Trans Internet Tecnol 2009.
[13] Nidhi Jain Kansal and Inderveer Chana (2012), "Cloud Load Balancing Techniques: A Step Towards Green Computing", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January.
[14] Soramichi Akiyama, Takahiro Hirofuchi, Ryousei Takano, Shinichi Honiden (2012), "MiyakoDori: A Memory Reusing Mechanism for Dynamic VM Consolidation", Fifth International Conference on Cloud Computing, IEEE 2012.
[15] Jyothi Sekhar, Getzi Jeba, S. Durga (2012)"A Survey on Energy Efficient Server Consolidation Through VM Live Migration", International Journal of Advances in Engineering & Technology, November..
[16] Rashmi. K. S, Suma. V and Vaidehi. M (2012), "Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud", Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June.
[17] Abdulaziz Aljabre (2012), "Cloud Computing for Increased Business Value", International Journal of Business and Social Science Vol. 3 No. 1; January.
[18] Abdullah Alshwaier, Ahmed Youssef and Ahmed Emam (2012), "A New Trend for E-Learning In KSA Using Educational Clouds", Advanced Computing: An International Journal (ACIJ), Vol.3, No.1, January.
[19] Cloud Computing for Dummies, Wiley Publishing, Inc.
[20] Rakhi k Raj and Getzi Jeba Leelipushpam. P (2012), "Live Virtual Machine Migration Techniques – A Survey", International Journal of Engineering Research and Technology, Volume 1 Issue 7, September.